ANAPHORA IN THE INTERLANGUAGE OF ENGLISH AND GREEK LEARNERS OF L2 SPANISH: A STUDY BASED ON THE CEDEL2 CORPUS

UNIVERSIDAD
DE GRANADA

Athanasios Georgopoulos

Supervised by: Cristóbal Lozano and Ana Díaz-Negrillo

GRANADA

2017

Στους γονείς μου

Editor: Universidad de Granada. Tesis Doctorales
Autor: Athanasios Georgopoulos
ISBN: 978-84-9163-655-7
URI: http://hdl.handle.net/10481/48765

ABSTRACT

It is a broadly accepted fact that the use and alternation of anaphoric forms in discourse (zero expressions, pronouns, nouns, etc.) is both syntactically and contextually constrained (Botley & McEnery, 2000; Huang, 2000; Nariyama, 2004; Rothman, 2009). It has also been demonstrated that L2 learners of several languages show persistent deficits concerning the interpretation and distribution of anaphoric subject expressions (for Spanish L2: Lozano, 2009b, 2016; Montrul & Rodríguez Louro, 2006). While research in this field has traditionally focused on anaphoric resolution (as opposed to production) and the bulk of findings rely on experimental data, there is an increasing number of researchers who point out the need of using corpora to test existing SLA hypotheses (Díaz-Negrillo & Thompson, 2013; Granger, 2012; Lozano & Mendikoetxea, 2013; Mendikoetxea, 2013; Myles, 2015). Additionally, most anaphora studies in Spanish L2 have examined the interlanguage of English-speaking learners, whose L1 differs from Spanish with respect to the gamut of referential expressions (English, contrary to Spanish, is a non-pro-drop language), whereas the few studies that focus on non-Anglophone learners are usually concerned with the interpretation of anaphoric pronouns (Kras, 2008; Lozano, 2002b, 2002c, 2008b, forthcoming). Overall, in Spanish L2, there is a very limited number of production-oriented studies on the interlanguage of learners with pro-drop L1 background such as Greek (Margaza & Bel, 2006).

This thesis aims to explore the anaphoric 3rd person subject usage in the interlanguage of English and Greek learners of L2 Spanish at various proficiency levels. In addition, this study aims to provide a general account regarding the factors that constrain referential choices in Spanish L1. The integrated theoretical approach adopted here draws on relevant proposals from theoretical linguistics, psycholinguistics, computational linguistics and corpus linguistics (Ariel, 1990; Arnold, 1998; Givón, 1983; Gundel, 2010; Kibrik, 2011; Lozano, 2016; Mitkov, 2002; Ryan, 2015). The empirical database of the investigation is CEDEL2 (Lozano, 2009a; Lozano & Mendikoetxea, 2013), a written corpus that contains production data from English and Greek learners of L2 Spanish. Additionally, CEDEL2 contains data from native speakers of Spanish as a control corpus. Crucially, all three subcorpora exhibit the same design principles. Hence, this is the first corpus-based study in Spanish L2 that compares three proficiency levels of two groups of learners (whose L1 differs with respect to the distribution of anaphoric subjects) against a native control group. The main purpose of this thesis is to test several L2

acquisition hypotheses, focusing on the role of crosslinguistic influence on discourse anaphora.

The XML annotator UAM CorpusTool (O'Donnell, 2009) was used for the analysis of the corpus data. A fine-grained tagset was designed and a purpose-oriented Interlanguage Annotation was performed (Lozano & Díaz-Negrillo, submitted) following the methodology used in previous corpus-based studies on L2 Spanish (Blackwell & Quesada, 2012; Gudmestad, House, & Geeslin, 2013; Lozano, 2009b, 2016). Learners of three different proficiency levels (intermediate, advanced and upper-advanced) from each L1 background (English and Greek) were examined and compared against the native Spanish control group. Results showed that although intermediate and advanced Greek-speaking learners of Spanish show some tendency to overuse unpragmatic anaphoric subjects, they do so in a significantly lower percentage than their English counterparts. Moreover, at the upper-advanced level, the Greek-speaking learners exhibit native-like performance, in contrast to the English-speaking learners, who are overexplicit even at the highest levels of proficiency. Cross-linguistic influence may account for these differences between the two learner groups. Greek-speaking learners seem to take advantage of the similarity between their L1 and Spanish with respect to the distribution of anaphoric subjects, whereas English-speaking learners seem to transfer their corresponding L1 properties. However, the fact that the intermediate Greek-speaking learners are also occasionally overexplicit is in line with two postulations that have been very recently put forward in the literature. First, overexplicitness may be a universal tendency at the intermediate levels of proficiency of L2 acquisition. Second, no single factor may successfully account for learners' deficits and only the consideration of multiple factors that operate simultaneously may fully account for the observed non-target L2 performance.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my greatest gratitude to my two supervisors, Professor Cristóbal Lozano and Professor Ana Díaz-Negrillo, for their constant support and advice during the development of this dissertation. Without their guidance and help, this study would never have come to be. During the last four years, Ana y Cristóbal have been the '24/7 support center' of this study, always there to answer my questions and to encourage me to move forward when this was most needed. In addition, I would like to specifically thank Cristóbal for granting me full access to all the data of the CEDEL2 corpus that had been previously collected by him and his research team.

I also owe a warm thank you to each one of the 173 Greek-speaking participants of the corpus who responded to my calls for participation and took the time to write essays in Spanish and complete my questionnaires and forms. Their anonymous contribution has provided a great amount of original data that, apart from their importance for this thesis, may be used for many scientific purposes in the future. I would also like to gratefully thank the people who helped me with the collection of the data: Associate Professor Dimitra Papadopoulou and Assistant Professor Alexandros Tantos who actively promoted my data-collection project among their students in the Department of Philology at the Aristotle University of Thessaloniki. Assistant Professor Angelica Alexopoulou who asked her students in the Department of Spanish Language of the University of Athens to participate. Finally, all my friends and colleagues who also asked other persons with knowledge of Spanish from their social circles to participate.

I am also grateful to my 'academic' friends Pandelis, Katerina, Gari, Mónica, María and Dimitris who were always available to drink a wine with me and talk about academic and scientific matters that only them -among all my friends- would understand. Finally, special thanks go to my friends Panagiotis, Stella, Aggeliki, Giorgos, Giagos and Alexandra for their psychological support whenever needed. I am grateful to each of you mentioned here and to all the people who have supported me during the last four years.

This dissertation is dedicated to my parents, Nikos and Ntina, for their unconditional support and love. Thank you with all my heart.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

AJT	Acceptability Judgment Task
AR	Anaphora Resolution
CA	Contrastive Analysis
CEDEL2	Corpus Escrito del Español como L2
CFC	Contrastive Focus Context
CIA	Contrastive Interlanguage Analysis
CMFT	Context-Matching Felicitousness Task
EA	Error Analysis
GJT	Grammaticality Judgment Test
IH	Interface Hypothesis
ILA	Interlanguage Annotation
IPH	Interpretability Hypothesis
LCR	Learner Corpus Research
MA	Moroccan Arabic
NP	Noun Phrase
NSP	Null Subject Pronoun
OPC	Overt Pronoun Constraint
PAS	Position of Antecedent Strategy
PPVH	Pragmatic Principles Violation Hypothesis
PRI	Potential Referential Interference
PVT	Picture Verification Task
RT	Reaction Time
SDRT	Segmented Discourse Representation Theory
SLA	Second Language Acquisition
SPRT	Self-Paced Reading Task
TC	Topic Continuity
TMA	Tense Mood Aspect of the verb
TS	Topic Shift
UA	Underspecification Account
UDH	Unidirectionality Hypothesis
UG	Universal Grammar
WCT	Written Contextualized Task

CHAPTER 1

1 INTRODUCTION

The main purpose of this thesis is to explore the acquisition of anaphoric subjects by English and Greek adult learners of L2 Spanish[1]. In addition, this study aims to provide a general account regarding discourse anaphora in Spanish L1. More specifically, the focus of interest is on the production of 3rd person anaphoric subject expressions. All the data of this study have been extracted from CEDEL2 (*Corpus Escrito del Español como L2*, Lozano, 2009a; Lozano & Mendikoetxea, 2013) which is a written Spanish L1/L2 corpus. The present corpus-based study aims to contribute in bringing together Second Language Acquisition (SLA) and Learner Corpus Research (LCR).

Etymologically, anaphora (originating from Ancient Greek: *αναφορά*) means repetition. The word is composed of the prefix *ανα* ("re") and the verb *φέρω* ("to bring", "to bear"). In Modern Greek, though, the primary meaning of *αναφορά* is reference. Historically, the study of anaphora/reference has been of major interest to scholars from several disciplinary fields. Especially in some areas of Philosophy, Linguistics and Psychology the concept of reference is fundamental (e.g. in Philosophy of Thought: Wettstein, 1984, in Computational Linguistics: Mitkov, 2002 and in Cognitive Psychology: Garnham & Oakhill, 1990). According to Sullivan (2006:420), "philosophical problems that turn on the notion of reference are more or less as old as philosophy". On the other hand, in formal linguistics and psycholinguistics, the study of reference is broadly labelled under the term 'anaphora resolution' (AR).

What is it that makes anaphora so interesting though? Before we provide an answer to this question, a working definition of the concept is in order. Lyons (1968:404), provides a fairly broad one: "Reference is the relationship which holds between words and things". This definition is similar (but not identical) to the current point of view in psycholinguistics, where it is broadly assumed that AR is "the mental association of real-world entities with referential linguistic expressions" (Jegerski, VanPatten, & Keating, 2011:483). Finally, a more strict textual definition is provided by Lozano

[1] The present study follows Myles (2015:310) in that the term L2 is used to refer to any language acquired after the native language has been acquired. Thus, no distinction is being made herein between L2, L3, L4, etc. nor between the acquisition of 'second' and 'foreign' languages. The acquisition process is assumed to be similar in all the aforementioned cases.

(2016:237) who argues that AR "relates to how an anaphoric expression (NP/overt pronoun/Ø pronoun) corefers with its antecedent in the discourse". Note here that the aforementioned definitions are representative of three different views upon a crucial matter regarding anaphora, namely: where does the anaphoric linguistic expression refer to? The answer to this question is precisely what makes the study of anaphora so complex and intriguing. The anaphoric linguistic expression (henceforth: anaphor) may refer to things in the real world (in the first definition), to representations of things in our minds (in the second definition) or to some other words in the discourse (in the third definition). An integrative approach is adopted in this study, insofar as all the above definitions are considered valid and complementary to each other. Throughout this dissertation, however, I will conventionally use the last one, as a working definition for anaphora. The reason for this methodological choice is simple: although the text is not a perfect reflection of the psycholinguistic processes taking place in the mind, it may certainly provide some insights, as previous literature on discourse analysis has revealed (Givón, 1983; Halliday & Hasan, 1976; *inter alia*). It is precisely in the text where the referent (whatever this might be) is being expressed in various positions, adopting different linguistic forms at different moments. I will argue together with Emmott (1997:62) that, for the moment, no one is able to directly observe in detail the precise psycholinguistic processes taking place into the human mind and that "the language is not direct proof of mental processes, but it can be assumed to give some indication of what is going on in the mind".

The lack of consensus regarding the nature of the referent (thing, representation or word) is just one of the features that make the study of anaphora such a complex and fascinating issue. If we turn our attention to the purely linguistic constituent of the relationship (the anaphoric expression), we observe that various linguistic forms may be used in order to refer to the same thing. In many languages, the anaphor may even be absent in discourse (as in the case of null subjects/objects). Consider the examples in Spanish below (the discourse context being about "Julia and her daughter"):

1) a. **Julia** la quiere
   "**Julia** loves her"
   b. **Ella** la quiere
   "**she** loves her"
   c. **Ø** la quiere
   "(**she**) loves her"

In example (1a), the noun phrase *Julia* could be assumed to refer to the particular mother of the specific linguistic/situational context (although it could also refer to any other Julia in the world). The pronoun *ella (*"she") in (1b) could refer either to the mother or the daughter (or to someone else). The null subject *Ø* in (1c) could also refer to all the above. Note, on the other hand, that all three anaphoric subject forms may refer to the exact same thing (whatever that might be). The analysis would become even more intricate if we considered the other anaphoric expression as well (the object pronoun "her"). Regarding the complexity of referential procedures, there is broad consensus in the literature with respect to the fact that anaphora is both syntactically and pragmatically regulated (Huang, 2000a; Rothman, 2009; Sorace, Serratrice, Filiaci, & Baldo, 2009; *inter alia*). On the other hand, the observation that the same anaphoric form may be used to refer to different things whereas different anaphoric forms may refer to the same thing led Fretheim & Gundel (1996:7) to indicate that "the fact that people actually manage to understand one another most of the time seems almost magical".

This study aims to contribute to a better understanding of reference by examining the problem of anaphoric distribution in discourse which, according to Huang (2000:151), refers to "how to account for the choice of a particular referential/anaphoric form at a particular point in discourse". Surprisingly, despite the enormous body of literature on anaphora, the vast majority of studies are exclusively concerned with the other "side of the coin" (a term used by Arnold, 1998:66), namely the interpretation of the anaphor from the part of the listener/reader. As we will see in Chapter 3, only a small number of anaphora studies focus on production data and no more than a handful of them are interested in L2 discourse (Blackwell & Quesada, 2012; Gudmestad, House, & Geeslin, 2013; Lozano, 2009b, 2016). Given the popularity of the subject and the complexity of the phenomenon, one might expect the SLA literature to be full of studies on the anaphoric production of L2 learners. There are two, complementary, reasons as to why this not the case. Firstly, the technical impediments regarding the study of discourse anaphora. Until very recently, researchers did not have the means to collect and analyse the amount of data which is needed for the study of real discourse (Díaz-Negrillo & Fernández Domínguez, 2006:85). Secondly, the predominant formal SLA approach in the study of anaphora has not been very enthusiastic about the idea of examining real discourse (Quesada & Blackwell, 2009:117).

On top of all the above considerations, it should be noted that the literature on language combinations which do not include the English language (either as L1 or as L2) is almost

inexistent in the domain of discourse anaphora. Regarding specifically the study of anaphora in Romance languages[2], most previous SLA literature has focused on the interpretation of Spanish and Italian anaphoric subjects by advanced English-speaking learners who have been found to show persistent deficits (Lozano, 2009b; Montrul & Rodríguez Louro, 2006; Sorace, 2000; *inter alia*). Regarding the non-target performance of L2 learners, a particularly influential account (the Interface Hypothesis (IH): Sorace, 2011; Sorace & Filiaci, 2006) understates the effect of transfer as the crucial source of deficits and promotes an alternative explanation on the basis of processing difficulties. The initial hypothesis predicts that "properties involving syntax and another cognitive domain may not be fully acquirable" (Sorace & Filiaci, 2006:340). Anaphoric interpretation and distribution is one of the privileged locus of interest of the IH, since the phenomenon is assumed to involve both syntax and discourse. Learners' non-target performance (mostly overproduction of overt subject pronouns) has been attributed to the processing difficulties arising from the complex interaction between syntactic and discursive factors that determine the selection of anaphoric forms. However, as already noted, the bulk of evidence supporting the hypothesis derives from English-speaking learners where transfer may not be a priori discarded (but see: Bel & García-Alcaraz, 2015; Lozano, forthcoming). Myles (2015:315) argues that "to fully understand the development of null subjects in Italian (…) SLA researchers need to compare how learners from both null-subject and non-null-subject languages acquire it". Additionally, the majority of the studies focus on anaphora interpretation (as opposed to production) and adopt experimental methodology (but see: Blackwell & Quesada, 2012; Gudmestad, House, & Geeslin, 2013; Lozano, 2009b, 2016). Given the existing consensus in the literature that anaphora is to large extent regulated by the information status of referents (Quesada, 2015:272), the isolated artificial sentences which are broadly employed in experiments may not constitute ideal data when the purpose is to account for the distribution of anaphoric forms in discourse. Finally, the examination of only one proficiency level (usually advanced or upper-advanced) does not allow to have a

[2] Note here that a commonly accepted assumption in the literature considers a basic distinction between pro-drop languages (where null subjects are allowed) and non-pro-drop languages (where null subjects are not allowed) (Chomsky, 1981; Jaeggli, 1982). Romance languages (such as Spanish and Italian) and Greek are considered prototypical examples of the first type whereas English is representative of the second type (Biberauer, Holberg, Roberts, & Sheehan, 2010:7).

complete view of the developmental procedure of the learners. All the above caveats have been specifically considered and treated in this study with the employment of LCR methodology (Granger, Gilquin, & Meunier, 2015).

A widely accepted definition of learner corpora is provided by Granger (2002:5): "Computer learner corpora are electronic collections of authentic FL/SL textual data according to explicit design criteria for a particular SLA/FLT purpose". Despite the fact that LCR methodology has been rarely used in Spanish L2 anaphora literature (with some exceptions: Blackwell & Quesada, 2012; Gudmestad, House, & Geeslin, 2013; Lozano, 2009, 2016), the role of learner corpora has been constantly increasing in other SLA research domains over the last two decades (Granger, Gilquin, & Meunier, 2015:1). Recently, several scholars have pointed out the need of a closer cooperation between corpus specialists and SLA researchers in order to compile well-designed large databases of L2 textual production (Díaz-Negrillo & Thompson, 2013; Lozano & Mendikoetxea, 2013; Myles, 2005, 2015). This will allow us "to test SLA theoretical constructs on the basis of learner corpus data" (Granger, 2012:13). At the same time, there is a need to surmount several drawbacks in LC that have been reported in the past by promoting the integration of some significant improvements, namely: the adaptation of strict corpus design criteria (Lozano & Mendikoetxea, 2013; Mendikoetxea, 2013), the incorporation of a wide range of languages and levels (Myles, 2015), the use of objective proficiency measures (Tono, 2004), the employment of sophisticated software tools for the analysis (Gries, 2012) and the open access to the corpus data (Granger, 2002). The present study is based on data extracted from CEDEL2, a corpus that fully incorporates the aforementioned improvements. Furthermore, the empirical database of this study allows for the implementation of a fine-grained and purpose-oriented Interlanguage Annotation (Lozano & Díaz-Negrillo, submitted). whereas the use of a sophisticated corpus software (UAM CorpusTool, O'Donnell, 2009) provides the means for the implementation of a meticulous Contrastive Interlanguage Analysis[3] (CIA). CIA is a term proposed by Granger (2004:134) to designate an LCR methodology that mainly focuses on

[3] It should be noted here that, although the term CIA was originally proposed by Granger (2004), similar methods (interlanguage comparisons) have been used in SLA research long before the existence of LCR (Bley-Vroman, 1983; Long & Porter, 1985; *inter alia*).

"comparisons of native and learner data and comparisons of different interlanguages to each other". Regarding the advantages of CIA methods, Rankin (2015:236) argues that:

> "(CIA) permits the identification of transfer by comparing L2 production from different L1 groups. It also identifies patterns of divergent production as well as over and underuse by employing comparisons with native-speaker production"

In this thesis, the main purpose of the analysis is to test specific hypotheses regarding the acquisition of anaphoric subjects by learners of L2 Spanish from different L1 backgrounds (English and Greek) at three proficiency levels (intermediate, advanced, upper-advanced). Comparable data from a native Spanish control group will serve as the benchmark for comparisons. In addition, the anaphoric production of native Spanish speakers will be analysed separately. In sum, the main novelty of the present study lies in the contrastive analysis performed between learners of Spanish (at three proficiency levels) with two L1 backgrounds that differ with respect to anaphora (Greek pro-drop / English non-pro-drop).

The remaining of this thesis is structured as follows:

Chapter 2 deals with the theoretical background of the present study. Different theoretical approaches and key aspects of discourse anaphora are presented in this chapter, which aims to provide an overview of the existing theoretical models on discourse anaphora.

Chapter 3 contains an overview of the previous research on the L2 acquisition of anaphoric subjects. The aim of this chapter is to present the results and the claims made in previous SLA literature regarding the main subject of this dissertation: the L2 acquisition of 3rd person anaphoric subjects.

Chapter 4 presents the research questions and the hypotheses of the present study, formulated on the basis of the findings, the claims and the unresolved issues in previous SLA research.

Chapter 5 deals with the methodology of this study. The corpus database, the participants, the annotation software and the annotation scheme are presented in this chapter.

In **Chapter 6** the results of this study are presented and briefly discussed. The first section deals with the overall distribution of anaphoric subjects in Spanish L1 and Spanish L2. The factors that constrain referential choices in Spanish L1 are presented in the second

section. In the third section, a CIA is performed between the L2 groups and the control group.

In **Chapter 7**, the results of this study are broadly discussed in light of the research questions and the hypotheses presented in Chapter 4.

Finally, in **Chapter 8**, the conclusions of this thesis are resumed. Additionally, some recommendations for future work are made on the basis of the limitations of the present study.

CHAPTER 2

2 ANAPHORA: A THEORETICAL OVERVIEW

This chapter examines the main theoretical views on anaphora in linguistics (theoretical, computational and psycholinguistics). Although many linguistic studies on anaphora either adopt a mixed theoretical approach or avoid entering into notional debates, it is crucial to understand the origin of some fundamental concepts on anaphora, namely the *anaphor* and the *antecedent*. Werth (1984:124) offers a rather broad definition of anaphora, according to which:

> "Anaphora is a semantic relationship between an entity (call it A) which may be linguistic or not, and another one (call it B), which has to be linguistic, such that in some text world B corresponds to A"

This theoretical definition shall be initially adopted in order to examine the fundamental approaches to anaphora and arrive at a more practical one which will be used during the analysis of the data. In the above definition, *B* corresponds to the *anaphor* and *A* to the *antecedent*.

2.1 Syntactic approaches: Binding Theory

Binding Theory was originally proposed by Chomsky (1980, 1981, 1986) under the 'Principles and Parameters' framework. Although it has undergone several modifications (see Büring, 2005 for an overview), all versions of the theory are almost exclusively concerned with sentential anaphora (as opposed to discourse anaphora). The three principles of the original version of Binding Theory may be resumed as follows[4]:

A An anaphor must be bound in its binding domain
B A pronoun must be free in its binding domain
C An R-expression must be free in its binding domain

More specifically, the focus of interest of Binding Theory is mainly on reflexives and reciprocals such as the ones in the following example:

[4] An extensive review and analysis of Binding Theory is out of the scope of this study (for more details see Haegeman, 1991).

2) a. Mary$_i$ loves **herself$_i$**.
b. The two sisters$_i$ hated **each other$_i$**.

According to Binding Theory, the anaphoric forms in the above example ("herself", "each other") are grammatically constrained (*c-commanded* in terms of the theory) by their respective antecedents ("Mary", "the prisoners"). The antecedent in this approach is considered to be in the text, in the same sentence as the anaphor. Although an extensive review of Binding Theory is out of the scope of this study, two important observations are in order here. First, as already noted, this approach does not take into account anaphora outside the sentential domain. Second, and most important, it considers exclusively the role of syntactic factors on anaphora (Gardelle, 2012:30). Consequently, Binding Theory cannot account for complex cases of anaphora (especially in L2 discourse) such as in the example (3) below, extracted from a text of the CEDEL2 corpus[5] (the text is about Antonio Banderas and belongs to a Greek-speaking learner of Spanish):

3) Él$_i$ tiene 54 años. Su$_i$ madre$_j$, Ana Banderas Gallego$_j$, era professora$_j$ y su$_i$ padre$_k$, José Domínguez$_k$, era de policía$_k$ en España. **Él$_i$** tiene un hermano$_l$ menor, Javier$_l$ (GR21_22_1_2_JUA)

"He$_i$ is 54 years old. His$_i$ mother$_j$, Ana Banderas Gallego$_j$, was a teacher$_j$ and his$_i$ father$_k$, José Domínguez$_k$, was a police officer$_k$ in Spain. **He$_i$** has a younger brother$_l$, Javier$_l$"

The anaphoric subject under consideration in example (3) corefers with the possessive pronouns in the second sentence. It also corefers with the anaphoric subject of the first sentence. All of them corefer with other previous and subsequent linguistic forms in the text as well. It is out of the scope of Binding Theory to account for the anaphoric pronoun

[5] All authentic corpus examples in this thesis are conventionally presented as follows: the relevant anaphoric subject expression is in bold, the coreferring linguistic forms are co-indexically marked (with letters: i, j, k, l etc.) and different referents are marked with different indices. Regarding the text ID (in parentheses): all texts in the dataset have been coded according to the participant's data as follows: the initial letters correspond to his/her L1 (ENG: English, GR: Greek, SPA: Spanish), the next four numbers separated by dashes correspond (in this order) to: proficiency score in the test (0-43), age, length of instruction and composition title number (1-12). Finally, the last letters correspond to the participant's initials. The original texts where examples belong to can be easily retrieved by entering the corresponding text ID (without the L1 identifier) in the online search interface of the CEDEL2 corpus: http://cedel2.learnercorpora.com/. The English translations of the examples are provided in quotation marks below the originals (Spanish null subjects are in parentheses in the English translations). The translations are intended to be understandable but, at the same time, as literal as possible so as to maintain the essence of the original text.

under question, since (a) the antecedent is found outside the sentential domain and (b) the antecedent does not syntactically constrain the anaphor. Note, additionally, that although the pronoun in the example is grammatically correct, it could be replaced by two equally correct linguistic forms, namely a noun phrase or a null subject. As a matter of fact, both of them would be pragmatically more appropriate than the pronoun in the above example (see section 2.4.3 for more details). Crucially, Binding Theory cannot account for pragmatic appropriateness, due to the fact that the reasons behind the selection of anaphoric forms in discourse may not be fully accounted for in terms of syntactic constrains.

It should be noted here that, despite the criticisms it has received from pragmatic approaches to anaphora (Huang, 2000a; Levinson, 1991), Binding Theory has been highly influential (mostly, but not exclusively, in generative approaches). This study, however, agrees with Huang (2000a:90) who argues that "a single or a few syntactic parameters/features/rankings may never be adequate to account for this complex phenomenon". Given that the present study focuses on anaphora in discourse (extrasentential), we will now turn our attention to the discourse-oriented approaches presented in the next section.

2.2 Discourse approaches: textual model and pragmatic accounts

Outside the purely syntactic account, anaphora in discourse has been broadly studied in the fields of discourse analysis, computational linguistics and cognitive psychology. Regarding the *anaphor*, there is a general consensus in the literature, insofar as it is broadly defined as a dependent linguistic form. However, a fundamental question which concerns every single approach to discourse anaphora arises: what is the *antecedent*? Regarding its definition, there are two different views in the literature. On one hand, the textual models consider the antecedent part of the text. On the other hand, the cognitive approaches assume that the antecedent is a mental representation of the referent. A brief overview of the two models is in order before we present the approach adopted in the present study (see section 2.3).

2.2.1 The 'referent in the text' approach

The textual models of discourse anaphora are based on the seminal work of Halliday & Hasan (1976) who focuses on the role of anaphors as cohesion devices. In words of the authors (p. vii):

> "Cohesive relations are relations between two or more elements in a text that are independent of the structure; for example between a personal pronoun and an antecedent proper name, such as *John…he*. A semantic relation of this kind may be set up either within a sentence or between sentences"

Under this textual approach, the anaphor and the antecedent are two segments of the text that corefer. They are both regarded as linguistic instantiations of the referent, whereas the anaphor depends on the antecedent for its interpretation (resolution). The 'referent in the text' approach (a term used by Emmott, 1997:199) has been highly influential in several linguistic research fields. This notion of anaphora is regularly employed in computational linguistics (Mitkov, 2002; Schmolz, 2015). Traditionally, it is also adopted in grammatical studies (Simpson & Weiner, 1989; Stirling & Huddleston, 2010) and commonly accepted in generative theoretical approaches (Reinhart, 1983). Finally, in formal approaches to anaphora in SLA (Al-Kasey & Pérez-Leroux, 1998; Belletti, Bennati, & Sorace, 2007; Liceras, 1989; Sorace & Filiaci, 2006) it is also assumed that the anaphor refers back to its antecedent, located in the previous sentence or clause.

A crucial observation is in order here: although most recent formal approaches unanimously recognize the role of discursive/pragmatic factors in anaphora (in contrast to the tenets of Binding Theory), they do not specifically examine the relevance of such factors due to methodological limitations: the examination of isolated artificial sentences, commonly employed in the formal literature, does not allow the consideration of discursive/pragmatic factors. This study will contribute to overcome the limitations of SLA sentence-based studies by examining anaphora in real discourse as done in recent learner corpus-based research (Blackwell & Quesada, 2012; Geeslin & Gudmestad, 2016; Gudmestad, House, & Geeslin, 2013; Leclercq & Lenart, 2013; Lozano, 2009b, 2016; Ryan, 2015). Although the textual model of anaphora shall be broadly adopted here, it is important to understand its strengths and weaknesses as they have been pointed out by the pragmatic and cognitive approaches examined in the next section.

2.2.2 The 'referent in the mind' approach

Outside theoretical and computational linguistics, the predominant view on anaphora is quite different. The 'referent in the mind' approach considers the antecedent to be a mental representation of the anaphor, rather than a segment of the text. Emmott (1997:198) points out that mental representations, which are ignored in the textual model, are taken for granted in psycholinguistic approaches to anaphora (Garnham & Cowles,

2006; Nicol & Swinney, 2003). Additionally, the textual model has received important criticism from scholars working under cognitive/pragmatic frameworks (Brown & Yule, 1983; Cornish, 2006). The main objections of such frameworks, regarding the view adopted in the textual approaches (see previous section), are summarized in Cornish (2010:233) as follows:

i. The antecedent is a dynamically construed mental representation, rather than a static textual linguistic element.
ii. The mental representation accrues properties as discourse evolves: both the anaphor and the textual antecedent may contribute to this, rather than simply corefer with each other.
iii. There may be no textual antecedent at all in some cases of situationally constrained anaphora.

The author does not deny, however, the role of co-text in anaphora as he notes that "the co-text is only one ingredient in the establishment of an anaphoric interpretation" (p.227). He proposes, instead, a distinction between the 'antecedent' (mental representation) and the 'antecedent trigger' (textual antecedent). In other words, the author recognizes the role of the textual antecedent in anaphora but focuses instead on the mental representation of the referent in order to account for anaphoric relations.

Despite the discrepancies between textual and cognitive models of discourse anaphora, a review of the relevant research reveals that the above dichotomy is not reflected as such in the literature. An integrated view is adopted in several studies, insofar as both approaches are implicitly assumed to be valid and complementary. Ariel (1990) uses the term 'antecedent' for both textual antecedents and mental representations. The seminal studies of Chafe (1976, 1980) do not mention mental representations (although they are implicitly taken for granted in his work). Givón (1983), under a purely functional approach, aims at measuring 'topicality' in discourse anaphora by means of textual features. The author acknowledges that "the text does not reveal the assumptions made by speakers or hearers as to topic identifiability in a direct way" (p.12). He argues, however, that it reveals "the grammatical, 'purely linguistic' devices used to code various topics/participants in the discourse" (p.13). Kleiber (1994:35), in line with the aforementioned authors, also defends a cognitive approach but emphasises the need to consider the textual dimension as well, in terms of linguistic criteria.

In line with the above, the present study will adopt an integrated approach, insofar as the traditional textual definition of discourse anaphora is employed whereas the cognitive aspects of the phenomenon are also taken into account. It is hard to imagine a well-rounded approach to discourse anaphora which is not based, at least in part, on textual data. On the other hand, no such account may be comprehensive enough if cognitive processes in terms of mental representations are disregarded. Blackwell (2003:259) highlights the need for such an integrated approach:

> "Traditionally, the factors known to be involved in anaphora have been assigned to specific linguistic 'domains', such as the 'cognitive', the 'syntactic', the 'pragmatic', and the 'semantic' domains. That so far, no single theory, approach, or set of principles has been able to explain the entire complex problem surrounding discourse anaphora, points to the need to seek a better explanation, furnished by an integrated theory, drawing on useful notions from the different domains cited above."

We are now in a position to provide a working definition of the anaphor and the antecedent, in connection with the focus and the purposes of this study. This task will be presented in the following section.

2.3 The present approach

This thesis focuses on the entire set of 3rd person anaphoric subject expressions (null, overt pronouns and noun phrases) in Spanish written discourse. Although grammatical persons are merged in many previous studies on Spanish anaphoric subjects (more on this in Chapter 3) the present study argues for the need to separately examine the 3rd person anaphoric forms. The reasons for focusing only on 3rd person anaphors have been repeatedly highlighted in the literature (Benveniste, 1971; Chafe, 1994; Fernández Soriano, 1999; Geeslin & Gudmestad, 2016; Lozano, 2009a) and may be summarized as follows:

i. 1st and 2nd person anaphors constantly refer to the participants of the speech act (deictic function). In that sense, some scholars argue that only the 3rd person is properly anaphoric (see Lozano, 2009b for an overview).
ii. 1st and 2nd person anaphors may be only pronominal, whereas 3rd person anaphors comprise noun phrases (NPs) as well. Rosengren (1974:28) argues that "first and second person expressions are not true pronouns at all since they do not substitute for a noun".

iii. 1st and 2nd person, in contrast to 3rd person anaphors, may never have competing referents whose eventual presence has been shown to be particularly relevant in anaphora (Ariel, 1990; Givón, 1983; Lozano, 2016).

iv. The acquisition of 1st and 2nd person anaphors, in contrast to 3rd person, may not be problematic for L2 learners. Lozano (2009b) found that 3rd person singular human anaphoric subjects are the locus of deficits in L2 Spanish, whereas the rest of the pronominal paradigm (1st and 2nd person, plus 3rd person inanimate) is fully acquirable.

The above differences between grammatical persons render their separate analysis more than necessary, insofar as they are not fully comparable regarding their anaphoric properties. The inclusion of the entire pronominal paradigm in the analysis (a common procedure in several studies on discourse anaphora) may severely skew the results.

Regarding the dataset of this study, it consists of written narrative texts based on two tasks: a film-retell task and a famous person's biography task. Given the above observations on the necessity of separately examining 3rd person anaphors, the focus of the present study on narrative discourse is justified from the fact that 3rd person referents "typically recur in narratives more consistently than in many other discourse types" (Kibrik, 2011:13). It should be noted, however, that there is no clear-cut distinction between genre taxonomies (also known as rhetorical modes) in real discourse. A narrative text may also contain some non-narrative passages and vice versa. This thesis, however, together with other similar studies (Gudmestad, House, & Geeslin, 2013; Lozano, 2009b, 2016; Quesada & Blackwell, 2009), assumes that 3rd person anaphors are more likely to be found in discourse genres which are expected to contain narrative passages. Additionally, we agree with Saunders (1999:13) who proposes that "the ideal level for the examination of the acquisition of subject expression is the narrative structure". It should also be noted that narrative has been traditionally the genre *par excellence* in discourse anaphora studies (Chafe, 1980; Chen & Lei, 2012; Clancy, 1980; Emmott, 1997; Flores-Ferrán, 2002; Givón, 1983; Kang, 2004; Muñoz, 2001; Serratrice, 2007b; Travis, 2007). Finally, as Rohde & Kehler (2014:913) point out, "undoubtedly, the most well-studied referential form in psycholinguistics is the third-person pronoun".

Before we proceed with a more technical specification of the current approach, some observations regarding the communication mode under study are in order. It might be argued that anaphora should be ideally studied in oral discourse instead of the written texts which are examined here. There is no doubt regarding the need for oral data in

corpus studies, as it has already been pointed out by some authors (Díaz-Negrillo & Thompson, 2013; Myles, 2005, 2007). However, there are several facts that justify the selection of written essays for this study. First, and foremost, the methodological limitations regarding the research population (more on this in Chapter 5). The need for big amounts of authentic data, in order to rigorously examine learner production, makes the analysis of spoken anaphoric discourse a time-consuming task requiring a large amount of human resources in terms of transcription and coding. Additionally, as Granger (2002:8) points out, "the notion of authenticity is somewhat problematic in the case of learner data" since most L2ers are not very likely to spontaneously narrate stories in Spanish in their everyday lives. Even if they do so, in strictly controlled classroom settings, it would be very hard to collect the amount of oral data needed for a corpus study on anaphora[6]. Learners do, however, write narrative essays that may be massively collected even outside the classroom. Finally, written L2 data have been extensively collected and used in the past, as in the case of the well-known ICLE corpus[7] (Granger, 2009; Granger, Hung, & Petch-Tyson, 2002). Therefore, we may safely argue that "written language is as reliable as spoken language to study interlanguage phenomena, as shown by the numerous publications that have used written learner corpora" (Lozano & Mendikoetxea, 2013:81).

Finally, the exclusive focus of the present study on written narrative texts is justified by the fact that discourse anaphora may be affected by discourse genre and mode. As Myles (2015:315) points out: "it is crucial to be aware of the differences between different types of communicative activities". In line with this, we argue for the need to control for both discourse genre and mode, as done in the present study that focuses exclusively on written narrative discourse. The reason for this is that the distribution of anaphoric forms is expected to vary (to some extent) in different discourse genres and modes: e.g. in a written narrative text, an oral conversation and an argumentative essay. All in all, the written narrations which comprise our dataset may be considered among the prototypical discourse types and were found to contain numerous referents and intricate anaphoric relations (see Chapters 5 and 6).

---

[6] Consider that only the Greek-speaking learners' sample which was specifically collected for this study consists of 173 participants and a total number of almost 80.000 words.

[7] https://www.uclouvain.be/en-cecl-icle.html

In sum, the present approach on 3rd person anaphoric subjects responds to "the need for a synthesis of different approaches to discourse anaphora" (Botley & McEnery, 2000:3), This study is based on textual models, whereas the functional/cognitive aspects of the phenomenon are not discarded. All 3rd person subject expressions are considered in this study with the focus of attention being on the three prototypical anaphoric subject forms in Spanish, namely: null, overt pronouns and NPs. Anaphoric subjects are defined by their coreferential relation with a previous linguistic item, namely the antecedent. This is theoretically considered as a mental representation of the referent which is textually instantiated in the written essays under study. Although each of the preceding coreferential forms may be considered to be a textual antecedent of the anaphor (see example (3) in section 2.1), only the one located exactly before the anaphor will be treated in the analysis (see section 5.5.2 for more details). The "closest antecedent" (Mitkov, 2002:27) approach allows to overcome the technical impediments that would arise from the simultaneous consideration of the entire set of coreferring expressions. Most importantly, it allows comparability both with formal and functional approaches on anaphora where it has been broadly assumed that the textual antecedent is the closest linguistic form that corefers with the anaphor. The theoretical basis of this corpus-based study draws on the discourse anaphora models of Givón (1983), Ariel (1988, 1990), Gundel, Hedberg, & Zacharski (1993) and Kibrik (2011). The aforementioned studies, together with other similar approaches, shall be reviewed in the following section.

2.4 Key aspects of discourse anaphora

In the last three decades, discourse anaphora has been the subject of numerous studies in several linguistic fields[8] (Abbott, 2010; Ariel, 1988, 1990; Arnold, 1998; Botley & McEnery, 2000; Chafe, 1976, 1980; Cornish, 1999; Emmott, 1997; Fox, 1987a; Fretheim & Gundel, 1996; Gaillat, 2016; Gibson & Pearlmutter, 2011; Givón, 1983; Gundel, Hedberg, & Zacharski, 1993; Hinds, 1977; Hofmann, 1989; Huang, 2000a; Kibrik, 2011; Mitkov, 2002; Schmolz, 2015). The recurring debate in most of the studies on discourse

[8] Theoretical research on discourse anaphora is not concerned with L2 acquisition and focuses only on the interpretation and production of anaphoric expressions by native speakers of several languages (English, Spanish, French, Russian, etc.).

anaphora revolves around the prominence of the referents involved in anaphoric processes[9]. Several terms have been used quasi-synonymously in the literature in order to account for the unanimously accepted observation that some referents, at a given point in discourse, enjoy a more privileged status than others (Arnold, 2010:188). The labels that have been employed for the identification of the notions involved in the aforementioned idea include, among others, the following: topicality, salience, prominence, givenness, focus, activation, accessibility and attention. Despite the nuances in the definition of these terms, the common assumption which drives their original conceptualization may be formulated like this: the more important (topical, salient, prominent etc.) the referent, the less explicit the referential means used to refer to it. As we have seen, this assumption already sets off with a challenging assignment regarding the definition of the *referent.* Additionally, there is another complication which arises from the difficulty to measure the prominence of discourse referents. The notion of *topic* has been extensively used in the literature in order to account for this. However, the following observation of Reinhart (1981:53) is still valid today: "despite the intensive attention that linguists of various schools have paid to the notion *topic*, there is no accepted definition of it". All in all, it is crucial to understand how the notion of topicality and other relevant terms have been employed in anaphora studies. This is the purpose of the next section.

2.4.1 The sentential topic

The term *topic*, also known as *ground*, has been traditionally employed in linguistics in a complementary binary oppositional relationship with the corresponding terms *comment* and *focus* (Chafe, 1976; Gundel, 1974; Hockett, 1958; Kuno, 1972; Lambrecht, 1994; Vallduvi, 1992; Van Dijk, 1977). The former constituents of this theoretical bipolar relation, namely the *topic*/*ground*, are usually defined as "old information" (or "what the sentence is about"). The latter constituents, namely the *comment*/*focus* have been used to denote "new information" (or "what is being said about the topic"). In a parallel way, the

[9] Given the lack of terminological consensus in the literature, the broad definition of the term *referent* adopted in the present study includes both the real-world entity and its mental representation. It further includes its textual instantiations, namely the *anaphor* and the *antecedent*.

terms *theme* and *rheme* have also been used to denote a similar correlation (Firbas, 1966; Kuno, 1972; Mathesius, 1975). Despite the indubitable usefulness of the above-mentioned theoretical constructs in information-structure accounts[10], their application to empirical studies on anaphora has been proven to be problematic. We may distinguish two approaches with respect to the application of topicality in anaphora literature. Formal studies usually adopt a dichotomous view where one single referent in each sentence is considered to be more prominent. In functional approaches, however, topicality is contemplated in terms of a continuum.

A commonly found assumption in formal studies on anaphora is that one referent in each sentence (usually the grammatical subject) is the topic. This assumption, however, may oversimplify the application of the complex notion of topicality and has given rise to several difficulties. To begin with, there is an important confusion in the literature regarding the definition and use of the terms *topic* and *focus*[11]. Quesada & Blackwell (2009:121), for example, affirm that "focus can be new information, but it can also be information already introduced into the discourse and thus both the topic and the focus of the utterance". Montrul (2004:128) states that "overt pronouns can be focus (new information, contrast) or topic, or old information". In the above examples the terms *topic* and *focus* are employed as synonymous to both new and old information[12]. These examples are simply indicative of the perplexing application of the term *topic* in anaphora literature. The root of the problem lies, however, in a widespread misconception regarding the notion of topicality in discourse anaphora. Given that several anaphora studies point to the seminal work of Reinhart (1981) for the definition of sentential topics,

---

[10] According to Lambrecht (1994:5), information structure refers to "that component of sentence grammar in which propositions as conceptual representations of states of affairs are paired with lexicogrammatical structures in accordance with the mental states of interlocutors who use and interpret these structures as units of information in given discourse contexts"

[11] Prince (1981:225) provides an interesting anecdote regarding this issue: "as added evidence of the gravity of the situation, let me mention that the Old/New Information Workshop held at Urbana, Summer 1978, was quickly and quite appropriately dubbed the "Mushy Information Workshop"".

[12] Arnold, Kaiser, Kahn, & Kim (2013:410) point out that "some researchers use the term **focus** or **focus of attention** to identify the element that is more salient in discourse. But this use of the term has more in common with the linguistic term 'topic' than 'focus'".

it is crucial to understand what the author actually states with respect to this. Consider the following example extracted from her work (p.56):

4) Max saw Rosa yesterday.

According to the author, in the above sentence, both referents (*Max*, *Rosa*) can be topics: "It is a crucial fact about sentence topics, that equivalent sentences may have different topics (even if they mention precisely the same referents)" (p.58). Note that the application of the 'old information' criterion for the identification of topics is of little usefulness here. After their first mention in discourse, all referents become old information. As Reinhart (1981:74) points out, "old information is not sufficient to explain how we identify the topic of a given sentence, e.g. how we chose between the two equally 'old' candidates for Topichood". On the other hand, the 'aboutness' criterion is purely subjective since topics are not grammatically marked[13]. In example (4), the sentence may be argued to be about Max or about Rosa. It may also be about both of them or about something else (e.g. the fact that he saw her yesterday and not any other day). All in all, there are two direct implications regarding the above observations. Firstly, as Slabakova, Kempchinsky, & Rothman (2012:340) point out, "topic and focus can only be defined with respect to some discourse". This is in line with the observation of Taboada & Wiesemann (2010:1817) who argue that "the definition of topic relies on context". In the same line, Reinhart (1981:56) concludes that "'topic of' is a pragmatic relation, relative to discourse", Secondly, "at the local sentential level, topic position is not fixed and it is even possible to affirm that there may be more than one topic in a sentence" (Alonso, 2006:21). In sum, it has never been suggested in the theoretical literature on topicality that the terms *topic/ground/theme* correspond to a single referent of a particular clause[14]. As Van Dijk (1977:53) notes, "the topic need not be simply identical with any

[13] The fact that in some Asian languages (e.g. Chinese and Japanese), there is indeed a grammatically marked 'topic' has further contributed to the general confusion regarding the term. As Chafe (1994:84) notes: "The term topic (…) can be perhaps most usefully applied to a different phenomenon that is characteristic of Asian languages but its contribution to an understanding of English has been far from clear". Therefore, this notion should not be confused with the sentential topic as described here.

[14] The sentential topic examined here should not be confused with the discourse topic (Van Dijk 1977). The definition of the latter, which accounts for larger units of discourse, is not immune to problems either.

established discourse referent". It may be concluded from the above observations that the objective identification of a single topical element per sentence in real discourse, if feasible at all, is not a straightforward task.

In order to overcome the above-mentioned limitations, a dynamic conceptualization of topicality is adopted in most cognitive and functional accounts of anaphora[15]. The term 'communicative dynamism', initially proposed by Firbas (1956, 1966), aims to account for the relative informational value that a linguistic element acquires in the development of the communication[16] (Firbas, 1992:105). Under this view, topicality is thought to be a gradual property rather than a categorical feature. In the same line, Givón (1983:16) proposes the "degree of topic accessibility" and states that "it is clear that at least in some respect we are dealing here with a scalar, graded continuum". This dynamic view of topicality is implicitly or explicitly assumed in several more recent anaphora studies as well (Arnold & Griffin, 2007; Kibrik, 1996; Stevenson, Knott, Oberlander, & McDonald, 2000). As Zulaica-Hernández (2016) points out: "In such a dynamic context, language users establish a ranking of topicality among discourse referents". It is crucial to realize that this gradient model is not entirely incompatible with syntactic accounts on anaphora where the grammatical subject of a sentence is usually assumed to be the topic. In the dynamic approach, however, the syntactic function is merely considered as one of the many factors that "can either boost or dampen a representation" (Arnold, 2010:196). In other words, the referent in subject position may eventually be more topical than other referents, notably in the total absence of context, as in the case of the artificial examples commonly employed in experimental studies. See the example below extracted from Sorace & Filiaci (2006:188):

5) Il portiere$_i$ saluta il postino$_j$ mentre **Ø$_i$/lui$_{i/j}$** apre la porta.

However, in contrast to the sentential topic, the discourse topic may be objectively identified in some cases (e.g. when the 'topic of discussion' is previously announced in the title of a text).

[15] Some early cognitive accounts, however, are more in line with the traditional binary approach. Chafe (1976:28), for example, argues that "it has not been demonstrated linguistically that given vs new is anything more than a discrete dichotomy".

[16] Firbas (1992:104) acknowledges that the term 'communicative dynamism' was initially suggested to him by his teacher, Professor Josef Vachek, in a private communication.

"The porter$_i$ greets the postman while **Ø$_i$/he$_{i/j}$** opens the door".

The analysis employed by the authors (and commonly used in the formal literature) considers the subject of the first clause ("the porter") as the topic of the sentence. Additionally, it is assumed that the null subject of the second clause may only corefer with the alleged topic ("the porter"). In real discourse, however, there is always some previous and following discourse (the linguistic context)[17]. Following the analysis adopted in the formal literature, it could be the case that "the postman" had been the subject/topic of the previous sentence. Then a more natural version of the above example in discourse would be:

6) Il portiere$_i$ lo$_j$ saluta mentre **Ø$_j$/lui$_{i/j}$** porta la posta.
"The porter$_i$ greets him$_j$ while **Ø$_j$/he$_{i/j}$** brings the mail".

In example (6), it can be argued that the object ("him", referring to "the postman") is more topical than the subject ("the porter") since topicality in real discourse depends on several discursive and pragmatic factors as well (previous discourse, world knowledge, etc.). According to this dynamic approach, only the interplay of both syntactic and discursive features may determine the grade of topicality of each referent at any given point in discourse. The main point of this analysis is that the issue regarding the sentential topic is better formulated in terms of a gradient scale where a referent is "more topic" or "less topic" instead of "THE topic". In this sense, topicality refers to the information status of a referent and is being understood as synonymous to accessibility/activation/givenness/salience/etc. It remains to be seen, however, how the information status of a referent may be assessed in discourse. The present study follows some influential proposals to this respect, which will be examined in the next section.

2.4.2 Models of discourse anaphora

Four influential and interrelated theoretical models of discourse anaphora will be examined in this section. The Topicality Model, proposed by Givón (1983), was the first to provide an account of anaphoric distribution in discourse, in terms of measurable properties of linguistic expressions. In a similar line of research, Ariel's Accessibility Theory (Ariel, 1988, 1990) aims at assessing the information status of referents in the

[17] The picture gets even more complex if we consider that in real discourse there is usually a situational context as well. In order to simplify the analysis, the situational context will not be considered here.

minds of language users. Very similarly, Gundel, Hedberg, & Zacharski (1993) drew on the aforementioned seminal studies and proposed the Givenness Hierarchy which seeks to determine the correlation between linguistic forms and the cognitive statuses of referents. Finally, the Activation Model, proposed by Kibrik (2011), shares a lot with the aforementioned studies, insofar as it encompasses cognitive and computational approaches by proposing a correlation between linguistic expressions and the status of referents in working memory. Apart from the examination of the aforementioned models, this overview will consider some psycholinguistic approaches that are broadly compatible with the traditional cognitive views on discourse anaphora.

2.4.2.1 Givón's Topicality Model

Givón (1983) was the first who explicitly criticized the "vague and mysterious" (p.5) definitions of topic in the literature[18] and attempted to establish a measurable correlation between the topicality of a particular referent and its linguistic representation in discourse. Under the assumption that topicality is a graded continuum, four factors were initially proposed by the author to account for the 'topic availability' of any referent in discourse[19], namely:

i. Length of absence from the register ('distance', measured in number of clauses)
ii. Potential interference from other topics[20] ('ambiguity', measured in number of interfering referents)
iii. Availability of semantic information ('world knowledge')
iv. Availability of thematic information ('previous discourse knowledge')

However, the author finally developed discourse measurements based only on 'distance' (operationalized as the number of clauses between two mentions of a referent) and 'ambiguity' (operationalized as the number of other potentially interfering referents). It

[18] "Vagueness is simply the common scientific practice of handling one's readers a blank check, with the tacit understanding - or at least hope - that future research will fill in the detail" (Givón, 1983:36)

[19] 'Topic availability' is defined as "the *degree of difficulty* that speakers/hearers may experience in *identifying* a topic in discourse" (Givón, 1983:11).

[20] Notice that, under the assumption that all referents are topics (to some degree), Givón employs the term 'topic' as synonymous to anaphor/referent.

should be noted that both aforementioned factors have been considered in more recent production-oriented studies on anaphora (Blackwell & Quesada, 2012; Lozano, 2016; *inter alia*). The other two factors, namely (iii) and (iv), which largely correspond to general world knowledge and previous discourse information respectively, were not taken into account in Givón's study due to the difficulty in quantifying their contribution[21]. It should be noted here that the present thesis is the first study on discourse anaphora which attempts to perform, to some extent, the quantification of 'world knowledge' and 'previous discourse information' (see section 5.5.2.9 for details).

In their discourse measurements, Givón and colleagues later added a third factor under the label 'persistence' (operationalized as the number of clauses in which the referent is still present after its mention in the measured clause). The author further proposed the following scale aimed to graphically represent a principle according to which "the more disruptive, surprising, discontinuous or hard to process a topic is, the more coding material must be assigned to it" (p.18).

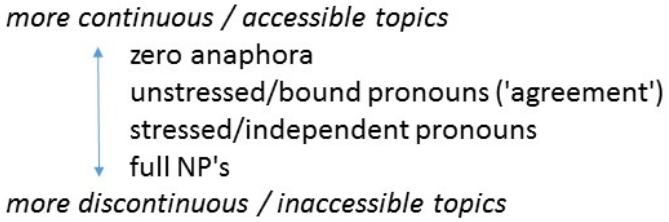


Figure 1. Ranking of the grammatical devices involved in coding topic accessibility (Givón, 1983:18)

The 'distance' factor of Givón's model, has been extensively employed in studies of discourse anaphora[22] (Arnold, 1998; Grüning & Kibrik, 2005; Gudmestad, House and Geeslin, 2013; Lozano, 2009b; *inter alia*) whereas the 'ambiguity' factor is also considered in several studies (Flores-Ferrán, 2002; Fox, 1987; Lozano, 2016; Sun & Givón, 1985; *inter alia*). Both factors have been empirically proven to correlate with the

[21] According to the author, other factors that were not included despite their undeniable importance are: "personality and memory of speakers and hearers, their specific life experience and the more subtle assumptions they make about each other and their respective abilities to identify referents" (Givón, 1983:12)

[22] It should be noted that prior to Givón (1983), Clancy (1980) also attempted to measure the effect of discourse factors (including 'distance' and 'ambiguity') to referential choices. Additionally, the role of distance has been considered in early psycholinguistic accounts to anaphora (Clark & Sengul, 1979).

choice of referential form. As Huang (2000b:153) notes, "there is compelling cross-linguistic evidence in support of the topic continuity or distance-interference model". However, in line with Arnold (1998:15) who notes that "Givón's measures of topicality (…) are too rough to accurately reflect the process of language comprehension and production", the present study argues for the need to adopt a more fine-grained approach (see chapter 5 for more details). Additionally, it has been argued that the predictions of the distance-interference model may be occasionally violated in two directions (Huang, 2000b:155-156). Firstly, in some, cases, "lexical NPs are used in discourse where distance is short and there is no interfering referent" (potential redundancy). Secondly, "reduced anaphoric expressions may be used over long distance" (potential ambiguity). In the first case, some stylistic or structural purposes (e.g. the beginning of a new paragraph in written discourse) may justify the technically redundant form, whereas in the second case, the role of context in terms of shared knowledge between the speaker and the addressee usually leaves no room for ambiguity. The present study aims to, at least partly, overcome these limitations by explicitly considering some of the factors which are not included in the distance-interference model (e.g. previous discourse and world knowledge, protagonisthood, etc.). Finally, it should be noted that Givón's textual model is, to some extent, comparable with other -purely cognitive- approaches (Ariel, 1988; Gundel, Hedberg, & Zacharski, 1993). Although his model focuses on the textual referent, rather than its mental representation, "Givón assumes that the text properties are associated with the cognitive status of entities" (Arnold, 1998:16). As Huang (2000b:157) points out: "the correlation between anaphoric encoding and topicality proposed by Givón can then be taken as a manifestation of the language user's cognitive status".

2.4.2.2 Ariel's Accessibility Theory

In a similar line with Givón's approach, Ariel (1988, 1990) proposed a comprehensive model in order to account for the status of the referent. Ariel's Accessibility Theory[23] is based on the assumption that "speakers choose their referring expressions by taking into consideration the degree of accessibility of the mental entity for the addressee (as best

[23] Ariel (1990:3) acknowledges that she adopted the term 'accessibility' which was initially proposed by Sperber & Wilson (1986).

they can assess it)" (Ariel, 1996:20). In other words, Ariel proposes that referential forms reflect the accessibility status of the referents in the mind of the addressees as assessed by the speakers. It may be argued that, at first sight, the empirical calculation and verification of accessibility in these terms seems a rather ambitious enterprise (Williams, 1988:349). Ariel, however, proposes a number of factors that contribute to the assumed Accessibility status of an antecedent[24] (Ariel, 1990:28):

a) Distance: The distance between the antecedent and the anaphor.
b) Competition: The number of competitors on the role of antecedent.
c) Saliency: The antecedent being a salient referent, mainly whether it is a topic or a non-topic.
d) Unity: The antecedent being within vs. without the same frame/world/point of view/segment or paragraph as the anaphor.

The first two factors ('distance' and 'competition') are very similar (if not identical) to the corresponding 'distance' and 'ambiguity' factors as proposed in the previously discussed Topicality Model (Givón, 1983). The 'saliency' factor, on the other hand, constitutes another example of the confusion regarding the notion of 'topic' in the literature (see also section 2.4.1). Ariel seems to assume that topicality (in terms of a discrete dichotomy) determines saliency which in turn is one of the factors that affect accessibility. This correlation implies that topicality, saliency and accessibility are not exactly the same thing. However, Ariel does not define what topicality and, hence, salience are[25]. Additionally, this assumption leads to some contradiction, insofar as accessibility (in contrast to topicality) is clearly considered to be a graded notion in her work. The last factor proposed by the author ('unity') reflects a commonly made assumption in the literature according to which the hierarchical structure of discourse

[24] Ariel (1990:17) argues that "it is the degree of Accessibility of the antecedent which is the crucial factor". This is assumed to reflect "the current status an antecedent is believed to have in memory". Afterwards, she empirically demonstrates that "pronouns favour a position where the antecedent occurs in the previous sentence" (p.18). It is obvious from the above quotations that Ariel employs the term 'antecedent' for both the textual and the mental representation of the referent.

[25] Ariel (1990:162) acknowledges, however, that the "antecedent Saliency is a function of both syntactic and non-syntactic aspects of the antecedent".

may drastically influence referential choices[26] (for an overview see Huang, 2000b:157-160). Ariel further assumes that the speaker is constantly monitoring the dynamically changing accessibility status of each entity (in the mind of the addressee) and "chooses a referring expression from among the list of expressions available in her particular language" (Ariel, 1996:21). The expressions are arranged on a scale, from high to low accessibility markers, in the following order:

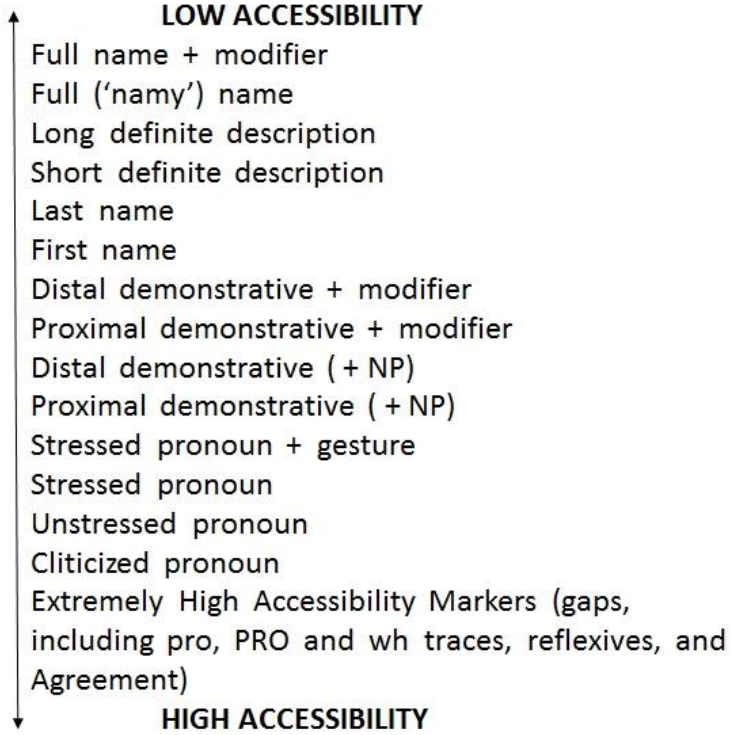


Figure 2. The Accessibility Marking Scale (Ariel 1990:73)

Ariel (1996:21) argues that the Accessibility Marking Scale is "by no means accidental". Three coding principles are assumed to predict the position of each referential form in the taxonomy. Firstly, 'informativity' accounts for the semantic content of the referential

[26] For some representative accounts on this respect see Hinds (1977), Hofmann (1989), Fox (1987a, 1987b) and Tomlin (1987). Despite the undeniable fact that discourse structure (in terms of episodes, events, themes, paragraphs etc.) is important in anaphora, the "hierarchy model approaches" (the term is proposed by Huang, 2000b:157) are unclear about the criteria which may conduce to an objective delimitation of discourse segments (except paragraphs which are graphically marked in a text). As Rasekh (1997:84) points out: "It is acknowledged that the characteristics of episode are weakly defined and are resistant to empirical analysis".

expression. It distinguishes, among others, zero from all explicit forms and pronouns from definite expressions. Secondly, 'rigidity', which determines how uniquely referring the expression is. As Filiaci (2010:65) points out, "this criterion is largely overlapping with the informativity criterion". Ariel (1996:21) argues, however, that "1st and 2nd person pronouns are more rigid than 3rd person pronouns" whereas all pronouns may be assumed to be equally 'informative'[27]. Lastly, 'attenuation' has to do with the 'phonological size' of the anaphoric expression. It distinguishes between forms that are equally informative but differ with respect to 'size' (e.g. stressed and unstressed pronouns). As Huang (2000a:255) points out, "there is cross-linguistic evidence in support of Accessibility theory in general and the scale in particular". Ariel has employed the same methodology as Givón (1983), namely discourse analysis, in order to demonstrate the validity of her claims. Despite the purely cognitive theoretical background of her work, Ariel seems to subscribe to the assumption that the text properties are associated with the cognitive status of referents. Therefore, and despite some technical differences regarding the nomenclature in the operationalization of discourse factors, Ariel and Givón are essentially making identical claims[28]. As already argued (see section 2.3), the theoretical approach adopted in the present study is also in line with the claim that referential linguistic expressions are textual instantiations of the cognitive status of referents.

2.4.2.3 Gundel's Givenness Hierarchy

A theoretical model of discourse anaphora which is often cited in the literature next to the ones of Givón (1983) and Ariel (1988, 1990) is the Givenness Hierarchy, proposed by Gundel, Hedberg, & Zacharski (1993)[29]. In line with the Accessibility Theory, as previously discussed, Gundel and colleagues make some crucial assumptions regarding referential expressions in discourse. Firstly, they argue that "different determiners and

---

[27] This is line with the present approach regarding the need for separation of 1st/2nd and 3rd person anaphors as described in section 2.3.

[28] As Eslami Rasekh (1997) points out: "Ariel (1988) acknowledges that Givón's theory on topic continuity, although the theoretical standing he ascribes to accessibility is not defined, is in the spirit of AT theory"

[29] The theoretical model of Gundel et al. (1993) is based, to some extent, on Chafe (1976, 1980) and Prince (1981) who systematically employed the term 'givenness' in their work (although generally adopting the traditional given/new dichotomous approach, where 'givenness' is merely synonymous to 'already mentioned').

pronominal forms conventionally signal different cognitive statuses (information about location in memory and attention state)" (Gundel et al., 1993:274). As Gundel (2010:151) makes clear, the Givenness Hierarchy is exclusively concerned with the specific kind of information encoded in referential expressions regarding "the addressee's assumed memory and attention state in relation to the intended referent (…) at the point just before the nominal form is encountered". Secondly, Gundel et al. (1993:276) propose that the different forms serve as processing signals to the addressee. Thirdly, it is being assumed that there are multiple linguistic and non-linguistic factors that determine how a referent comes to have a given cognitive status (Gundel, 2010:153) All the aforementioned assumptions are totally in line with the Accessibility Theory (Ariel 1988, 1990). Most importantly, Gundel and colleagues also proposed a referential hierarchy[30], aimed to graphically represent the correspondence between referential forms and cognitive statuses:

THE GIVENNESS HIERARCHY:

in focus	> activated	> familiar	> uniquely identifiable	> referential	> type identifiable
{*it*}	{*that*, *this*, *this* N}	{*that* N}	{*the* N}	{indefinite *this* N}	{*a* N}

Figure 3. The Givenness Hierarchy (Gundel, Hedberg, & Zacharski, 1993:275)

According to the authors, the 'in focus' status accounts for referents that are the current center of attention, as in the example below:

7) My neighbor's bull mastiff bit a girl on a bike. **It'**s the same dog that bit Mary Ben last summer.

The 'activated' status accounts for referents that are represented in short-term memory, for example:

8) I couldn't sleep last night. **That** kept me awake.

On the other hand, the 'familiar' status aims to account for referents that are represented in long-term memory, such as in the example below:

9) I couldn't sleep last night. **That dog** (next door) kept me awake.

[30] The authors examined five languages (English, Spanish, Chinese, Japanese and Russian) and proposed five different hierarchies according to the stock of anaphoric expressions in each language.

Regarding the 'uniquely identifiable' status, this is met when the addressee can identify the speaker's intended referent on the basis of the nominal anaphor alone. For example:

10) I couldn't sleep last night. **The dog** (next door) kept me awake.

On the other hand, the 'referential' status corresponds to cases where the speaker intends to refer to a particular object or objects. See the example below:

11) I couldn't sleep last night. **This dog** (next door) kept me awake.

Finally, the 'type identifiable' status accounts for mental representations that the addressee is able to access on the basis of the type of object described by the referential expression:

12) 1 couldn't sleep last night. **A dog** (next door) kept me awake.

Despite the similarities in the theoretical approaches of Ariel (1988, 1990) and Gundel et al. (1993), there are also some important differences between the corresponding models of Accessibility and Givenness. First and foremost, Gundel and colleagues do not provide any measurable factors which may be employed in order to examine the cognitive status of a referent. As Botley & McEnery (2000:10) point out, "unlike Ariel, Gundel does not give any specific metrics, such as textual distance, to allow us to measure the extent to which particular anaphoric expressions reflect particular cognitive statuses". Additionally, the Givenness Hierarchy is not an one-to-one mapping between forms and statuses (Gundel, 2010:155). Any cognitive status, as represented in Figure 3, entails all other lower statuses. What is 'in focus' is also assumed to be 'activated', 'familiar' etc. According to Gundel et al. (1993:294), "a particular form can often be replaced by forms which require a lower status". All in all, the Givenness Hierarchy was not developed to make predictions regarding a direct one-to-one mapping between the cognitive status of a referent and the anaphoric expressions employed to encode such status (Gundel, 2010:159). Consequently, unlike the models of Givón (1983) and Ariel (1988, 1990), the Givenness Hierarchy may not be objectively applied to direct measurements of 'givenness' in real discourse production. It may, however, provide valuable insights regarding the interpretation of idiosyncratic cases of discourse anaphora.

2.4.2.4 Kibrik's Activation Model

Recently, Kibrik (2011) put forward a discourse anaphora model which has a lot in common with the previously discussed proposals. Based on several cognitive-calculative

studies carried out by the author and colleagues[31] (Grüning & Kibrik, 2005; Kibrik, 1996, 2000, 2001; Loukachevitch, Khudyakova, Kibrik, Dobrov, & Linnik, 2011), the proposed model of discourse anaphora is grounded on the following three assumptions[32] (Kibrik, 2011:61):

i. Referential choice is immediately influenced by the referent's current degree of activation in the working memory of the speaker[33].
ii. The referent's current degree of activation depends on a wide range of factors stemming from the discourse context and the referent's internal properties.
iii. Each activation factor contributes a share to the referent's activation which can be numerically estimated.

The author has made several proposals over the last two decades regarding a wide range of factors which may contribute to the activation of a referent and, thus, may account for the choice of a referential expression at a particular point in discourse. The latest and more comprehensive model includes the following factors[34] (Kibrik, Khudyakova, Dobrov, Linnik, & Zalmanov, 2016:6):

[31] 'Cognitive-calculative' is a term proposed by Kibrik to describe his "arithmetical model that calculates referents' activation at any given point and thus accounts for any instance of referential device selection in the sample discourse" (Kibrik, 1996:255).

[32] Note that in this model there is an expressed preference for the term 'referential choice' over 'anaphora' (Grüning & Kibrik, 2005:2).

[33] Note that the notion of 'activation in the working memory' has been traditionally employed in psycholinguistics with similar meaning (see Arnold, 1998:57-65 for an overview).

[34] A detailed explanation of the factors included in Kibrik's Activation Model is out of the scope of this study. The reader is referred to the work of the author for more details (Kibrik, 2011).

(1) Referent's factors
• Animacy: animate, inanimate, collective (for such entities as organizations)
• Gender (for animate referents only): masculine, feminine, mixed (for groups of people with various or unspecified gender)
• Person: 1, 2, 3
• Number: singular, plural
• Protagonism: numeric value
(2) Anaphor's factors
• Ordinal number of referent mention in the referential chain: integer
• Type of phrase: noun phrase, prepositional phrase
• Grammatical role: subject, direct object, indirect object, oblique (with preposition), attribute, 's-genitive, of-genitive, postpositive specification
(3) Antecedent's factors
• Type of phrase (values same as in the section "Anaphor's factors")
• Grammatical role (values same as in the section "Anaphor's factors")
• Referential form:
◦ pronoun: personal, possessive, demonstrative, relative, zero
◦ description: a-description, the-description, bare description, demonstrative description, possessive description
◦ attributive
◦ numeral
◦ proper name: first, last, first and last, initials and last, non-person, acronym
◦ Antecedent length, in words: integer
(4) Distances between anaphor and antecedent
• Distance in words: integer
• Distance in all markables: integer
• Number of markables in chain from the anaphor back to the nearest full NP antecedent: integer
• Linear distance in EDUs: integer
• Rhetorical distance (RhD) in elementary discourse units: integer
• Distance in sentences: integer
• Distance in paragraphs: integer

Figure 4. Candidate factors of referential choice in Kibrik's Activation Model (Kibrik et al., 2016:6)

As can be seen in Figure 4, Kibrik's cognitive multi-factorial model builds on the previously examined proposals of Ariel (1990), Givón (1983) and Gundel et al. (1993). The author tested the model in L1 English and L1 Russian narrative texts through the implementation of sophisticated machine learning algorithms. His purpose was to predict the referential choice of the writer at any given point in discourse, based on the activation score of the referent as obtained by calculating the contribution of each individual factor. It should be noted that Kibrik's approach differs from the previously discussed models (Topicality, Accessibility and Givenness) in two important theoretical aspects. Firstly, it is exclusively speaker-oriented, insofar as no assumptions are made here regarding the mental state of the addressee. Recall that in previous accounts it is the status of the referent in the mind of the latter (as assessed by the former) that is aimed to be assessed. Secondly, the activation of each referent in Kibrik's model is calculated independently of the presence of other referents (as opposed to Topicality and Accessibility which depend on 'ambiguity' and 'competition' respectively). Kibrik presumes that a precise activation score may be calculated for each referent in discourse independently of the presence of

competing referents. Finally, Kibrik's model incorporates the factors proposed in the previous models in a more fine-grained manner (e.g. there are seven measurements of distance instead of one) and considers other factors as well (e.g. 'protagonism' of the referent). Despite these differences, Kibrik's cognitive-calculative proposal is clearly in line with the principal assumptions made in the previously discussed models of discourse anaphora.

2.4.2.5 Complementary approaches

The theoretical models discussed in the previous sections are also broadly compatible with several psycholinguistic approaches to anaphora (Almor, 2000; Arnold, 2015; Arnold, Kaiser, Kahn, & Kim, 2013; Gernsbacher, 1989, 1990; Kaiser & Trueswell, 2008; Sanford & Garrod, 1981). It should be noted, however, that most psycholinguistic studies focus on the processing/interpretation (as opposed to the production) of anaphoric expressions. Hence, real discourse data are rarely considered. In psycholinguistics (and in experimental studies in general) it is often assumed that "comprehension and production are merely two sides of the communicative coin" and that "the participants will develop similar mental representations about the shared discourse events and referents" (Arnold, 1998:66). Therefore, the information status of a referent at any given point in discourse is considered to be one and only one, in terms of the *common ground* shared by the speaker and the addressee. This allows studies on interpretation of anaphora to make claims about anaphoric production as well. The present study does not share this view and argues for a separate analysis of anaphoric production and interpretation, since the simultaneous consideration of the two processes may exponentially increase the complexity of the study of the phenomenon.

In one of the very few production-oriented studies in psycholinguistics, Arnold (1998) examined the effect of five discourse factors (Recency, Subjecthood, Focus, Parallelism, Goal Status) in written data from three languages[35] (English, Spanish and Mapudungun). The author assumed that the activation of referents in the working memory of the

[35] Recency refers to the 'distance' factor, as in the previously discussed models. Subjecthood refers to the referent being or not in subject position in the previous sentence. Focus, in this study, is defined in terms of left dislocation in cleft sentences. Parallelism accounts for the referents that have been last mentioned in the same grammatical role as the current referential expression. Goal Status is the property of a referent of having been last mentioned as the goal argument of a verb.

reader/listener correlates with the aforementioned linguistic factors. Hence, Arnold performed text analyses and demonstrated that referents which are less activated are more likely to be referred to with more specific anaphors. Leaving aside the existing difference of opinions in the literature regarding which mental representation should be examined (whether the speaker's or the addressee's or the addressee's as monitored by the speaker), Arnold's proposal is in line with the anaphora models proposed by Givón, Ariel, Gundel and Kibrik (see previous sections). The same is true for the studies of Kaiser and colleagues (Kaiser, 2003; Kaiser, Runner, Sussman, & Tanenhaus, 2009; Kaiser & Trueswell, 2008) who proposed the 'Form Specific Multiple Constraint'. Their approach is fully in line with the idea that anaphora in influenced by multiple factors. Additionally, Kaiser & Trueswell (2008:742) argue that "not all referential forms within a single language are sensitive to the same salience-influencing factors to the same degree".

Other psycholinguistic approaches have focused exclusively on the role of semantic factors in the interpretation of referential expressions. For example, 'implicit causality' verbs (Caramazza, Grober, Garvey, & Yates, 1977; Garvey & Caramazza, 1974; Goikoetxea, Pascual, & Acha, 2008; Hartshorne, Sudo, & Uruwashi, 2013; McKoon, Greene, & Ratcliff, 1993) have been assumed "to create a bias to re-mention the causally implicated referent" (Rohde & Kehler, 2014:918). In line with this, several connectives (Stevenson, Crawley, & Kleinman, 1994; Stevenson, Knott, Oberlander, & McDonald, 2000) have also been found to "project their own focusing preferences" (Miltsakaki, 2007:91). Note, however, that these approaches may only account for the interpretation of particular cases of discourse anaphora where the specific 'causality' verbs and/or connectives are involved.

On the other hand, in computational linguistics, anaphora has traditionally been a central issue (in automatic translation, information extraction, etc.). The bulk of relevant literature is concerned with natural language processing and the automatic resolution of anaphors which is a crucial issue whenever "text understanding is required or desired" (Schmolz, 2015:209). Interestingly, several computational approaches to discourse anaphora are also broadly compatible with the examined functional and cognitive models. Sidner (1981, 1983), for example, argues that "many constraints -syntactic, semantic, and pragmatic- affect the choice of specification for pronouns" (Sidner, 1981:229) and proposes the incorporation of cognitive notions in computational approaches to anaphora. Biber, Conrad, & Reppen (1998) and Botley (1999) have extensively examined referring expressions in written and spoken discourse in terms of information status as encoded by

multiple factors in texts ('distance' factors among others). The seminal study of Mitkov (2002) argues for the interaction of various factors which should be incorporated in any anaphora resolution algorithm. Finally, Botley & McEnery (2000) perform an extensive overview of computational approaches to discourse anaphora without neglecting cognitive models. They conclude that "a comprehensive general approach to anaphora resolution must take into account cognitive aspects, and aspects of discourse structure, as well as syntax and semantics" (p.11). In words of Mitkov (1994:1170):

> "Given the complexity of the problem, we think that to secure a comparatively successful handling of anaphora resolution one should adhere to the following principles: 1) restriction to a domain (sublanguage) rather than focus on a particular natural language as a whole; 2) maximal use of linguistic information integrating it into a unified architecture by means of partial theories."

The theoretical background of the integrative approach adopted in this thesis draws on the models proposed by Givón (1983), Ariel (1988, 1990), Gundel et al. (1993) and Kibrik (2011). It is further compatible with the predominant view in psycholinguistic approaches on discourse anaphora (Arnold, 1998, 2003). Finally, it incorporates methods which have been broadly employed in relevant studies in the field of computational linguistics (Biber, Conrad, & Reppen, 1998; Mitkov, 2002) and LCR (Blackwell & Quesada, 2012; Gudmestad, House, & Geeslin, 2013; Lozano, 2009b, 2016; Ryan, 2015).

However, given the focus of this study on learner discourse, a crucial observation is in order here. The bulk of theoretical models of reference are concerned with the study of anaphora in native languages. On the other hand, the field of anaphora in SLA is dominated by formal/syntactic approaches and the bulk of evidence derive from experimental comprehension-oriented studies (see Chapter 3 for an overview). One could argue that the proposed theoretical models may straightforwardly account for learner discourse but this might not be the case. The reason for this is not far to seek, given the existing consensus regarding the intricacy of the phenomenon. To begin with, the fact that each language has its own repertoire of referring expressions adds a considerable complexity layer to the study of anaphora in SLA. The corresponding correlations between forms and information statuses may be quite different across languages. Additionally, as it has been suggested (Arnold, 2010; Arnold, Kaiser, Kahn, & Kim, 2013; Sorace, 2006a), processing factors may also affect the production and interpretation of anaphors. Even if all others things were equal, this adds an extra burden for L2ers which could lead to variable linguistic behaviour with respect to referential production

and interpretation. Given the predominantly pragmatic nature of the phenomenon, reflected in the fact that all anaphors at any given point in discourse are grammatically correct, it would be reasonable to wonder how any potentially variable linguistic behaviour may be observed in real texts. The next section deals with this arduous task.

2.4.3 Ambiguity and redundancy

With respect to anaphora in SLA, there is a large number of recent studies which report that L2 learners of several languages tend to be 'overexplicit' (Chini, 2005; Gudmestad & Geeslin, 2010; Henriëtte Hendriks, 2003; Kang, 2004; Leclercq & Lenart, 2013; Lozano, 2016; Margaza & Bel, 2006; Ryan, 2015; Sorace & Filiaci, 2006). At the same time, in the opposite direction, some SLA studies also report that L2 learners are occasionally 'underexplicit' (Lozano, 2009b; Montrul & Rodríguez Louro, 2006). The terms 'overspecification' and 'underspecification' have also been employed in par with 'redundancy' and 'ambiguity' for the description of the above-mentioned linguistic behaviours. The same phenomenon has concerned anaphora studies on adult bilingualism (Serratrice, 2007a; Serratrice, Sorace, & Paoli, 2004; Shin & Montes-Alcalá, 2014), child bilingualism (Andreou, Knopp, Bongartz, & Tsimpli, 2015; Serratrice, 2007b; Shin, 2012; Shin & Cairns, 2012; Sorace, Serratrice, Filiaci, & Baldo, 2009) and L1 attrition (Flores-Ferrán, 2004; Montrul, 2004a; Sorace, 2004; Tsimpli, Sorace, Heycock, & Filiaci, 2004). Finally, 'redundancy' and 'ambiguity' have been reported for the native control groups in several studies as well (Abreu, 2009; Alonso-Ovalle, Fernández-Solera, Frazier, & Clifton, 2002; Bel, Perera, & Salas, 2010; Fukumura & van Gompel, 2015; P. Hendriks, Koster, & Hoeks, 2014; Lozano, 2016; Perales & Portillo, 2007). It should be noted, however, that in the majority of these studies 'redundancy' and 'ambiguity' are defined in terms of differences between groups in the acceptability of anaphors in isolated sentences (acceptability judgment tests)[36]. More specifically, in the experiment-based SLA research, the learner groups are assumed to be 'redundant' or 'ambiguous' if they judge as (non-) acceptable the use of specific referential expressions in the same artificial sentences that the native control group accepts/rejects. Although in real texts the picture is more complex than that, an objective definition of 'redundancy' and 'ambiguity' in LCR is lacking (with the exception of Ryan, 2015). As Polio (1995:356) points out, given

[36] For the limitations involved in the use of AJT in SLA see also Sorace (2006b).

that referential choices are influenced by the surrounding discourse, "it is difficult even for a native speaker to state when a zero pronoun is used correctly or incorrectly". Without a technically-defined operational definition of 'redundancy' and 'ambiguity' it is hard to determine the extent of target-deviant performance in L2 discourse.

Regarding ambiguity in real language, Arnold (2010:188) argues that all references are ambiguous: "Even a very specific reference like Jennifer Arnold can refer to more than one person (Google phonebook alone lists over 100)". It is obvious that this observation regarding proper names is also valid, and to a much greater degree, in the case of pronominal and elliptical references (null subjects for example). At the same time, however, Otheguy & Zentella (2012:147) argue that "actual discourse is seldom ambiguous". The authors cite several studies that confirm this idea (Avila-Jiménez, 1995; Bentivoglio, 1987; Travis, 2005). This is also in line with Chafe (1990:315), as quoted in Kibrik (2011:65), who suggests that "ambiguity exists primarily in the imagination of 'exocultural' linguists, while for real speakers familiarity and context are likely to remove most problems of keeping third-person referents straight". In other words, whereas every referential expression is potentially ambiguous, almost nothing is really ambiguous in authentic discourse due to the contribution of context[37]. By contraposition, one could argue that almost every anaphor is technically redundant in real language. This has been empirically confirmed by Eslami Rasekh (1997) who removed several referential expressions (pronouns and noun phrases) from English newspaper articles and asked native speakers to identify and fill in the anaphoric gaps. Interestingly, the participants of this study managed to correctly identify almost all the referents of the empty anaphoric slots[38]. The author concluded that "inaccessible entities are not easily identifiable, but they cannot be said to be unidentifiable in the texts we have been observing" (p.166). It would be logical to assume that referential expressions which may be removed without causing ambiguity are unnecessary. Consequently, this would render redundant the greatest part of the anaphors in real discourse.

[37] This is also confirmed by the data of the present study. Only 8 'insolvably ambiguous' anaphoric subjects (accounting for the 0.3% of the total number of items) were found during the analysis of the texts (see 6.1).

[38] "The subjects could not successfully identify the referents of a few of the NP slots because the intended referent was not revealed by context (5.5% of the cases of identification)" (Eslami Rasekh, 1997:198)

Given the above observations, two crucial questions arise. First, how can we objectively operationalize and account for 'redundancy' and 'ambiguity' in discourse? And second, how can we ensure a valid comparability between native speakers and L2 learners with respect to this? An interesting proposal regarding this issue was recently put forward by Ryan (2015). The author argues for the operationalization of 'redundancy' in terms of accessibility (Ariel, 1990) "in order to confirm that systematic overexplicitness does indeed characterize L2 speech" (p.829). Based on the calculative model of Toole (1996), the author numerically assesses the degree of accessibility of each referent and examines the corresponding referential expressions used by native speakers and learners of L2 English for each degree. Ryan claims that this operationalization "appears not to have been attempted in previous L2 research" (p.829). In line with this particularly novel approach, the concepts of 'redundancy' and 'ambiguity' have been operationalized in this thesis (see section 5.5.3 for details) with respect to the presence of the anaphor in objectively defined discourse patterns (in terms of syntactic and discursive factors). This is the first study that provides a technical definition of 'redundancy' and 'ambiguity' in terms of the use of referential expressions in objectively defined discourse patterns. This allows valid comparisons to be made between groups regarding 'overexplicitness' and 'underexplicitness'. For the remaining of this study, the aforementioned terms are used synonymously to 'technical redundancy' and 'technical ambiguity' in that order.

2.5 Summary of the theoretical overview

In sum, the theoretical overview of discourse anaphora presented in this chapter has examined the predominant views in the relevant literature and some general consensus has been revealed regarding the crucial role of information status. This role is consistently highlighted in terms of the topicality/accessibility/givenness/activation of referents in discourse anaphora. The existing consensus regards the following facts:

i. Discourse anaphora is accounted for in terms of the information statuses of the referents: it is crucial to determine, though, how this statuses relate to the mental representations of speakers and/or addressees and to what extent they may be represented and assessed in tangible linguistic data.
ii. Information status changes dynamically as discourse progresses depending on multiple factors (both syntactic and pragmatic): the nature and individual contribution of these factors, though, needs to be clarified.

iii. Information status concerns a graded phenomenon rather than a discrete dichotomy: there is no consensus, however, regarding the level of granularity and the specific grades of the corresponding referential hierarchies.

iv. The different echelons of the referential hierarchies correspond to different referential expressions: it is not clear, though, whether a direct one-to-one mapping between topicality/accessibility/givenness/activation grades and linguistic forms is feasible[39].

v. The correlations between information statuses and linguistic referential expressions may differ across languages. This stems from the fact that each language has its own repertoire of anaphoric expressions.

The present study builds on the above consensual observations and aims to perform a textual analysis (as described in section 2.3) based on real discourse data (coming from native Spanish speakers and learners of L2 Spanish). The main purpose of this study is to test several claims made in SLA literature regarding the L2 acquisition of anaphoric subjects, focusing on the entire range of 3rd person anaphoric forms in Spanish. A broad review of previous research on this matter will be performed in the next chapter.

[39] As Huang (2000b:163) notes, we are dealing here with an inherently imperfect correlation: "While the anaphoric coding contrast, for example, is in principle a matter of yes or no, activation is in principle a matter of more or less". In other words, there are not enough grammatical devices to represent the infinite number of discourse statuses.

CHAPTER 3

3 ANAPHORIC SUBJECTS IN SLA

Following the distinction made by Quesada (2015), the studies concerned with the acquisition of anaphoric subject expressions in a second language may be broadly divided into two main groups, according to the adopted theoretical line and methodology. On the one hand, formal/generative approaches are usually based on experimental methodology and seek to determine the underlying factors involved in the acquisition of anaphoric subjects by L2 learners. These approaches, which have dominated the field of anaphora in SLA in the last three decades, are mostly concerned with the interpretation (resolution) of anaphora from the part of the listener/reader. On the other hand, there is a handful of more recent discourse-oriented studies which, by nature, focus on the production (as opposed to the interpretation) of anaphoric subject expressions in real discourse. The latter approaches usually adopt either a variationist or a pragmatic theoretical view.

Regarding anaphoric subjects in Spanish L1/L2, Quesada (2015) offers a particularly extensive overview of the relevant literature. After examining the bulk of the studies and despite her previous categorization of studies between formal and functional approaches, the author concludes that "theoretical approaches like language itself, are never categorical in nature but can and do share characteristics with each other" (p.18). In line with this, the present dissertation takes a production-oriented, corpus-based approach and aims to test theoretical claims emanating mainly from formal models of anaphora. In line with the proposal of Quesada (2015:19), one of the central purposes of this thesis is to suggest avenues for future collaborative, cross-disciplinary research. The first section of this chapter deals with the theoretical assumption that the grammars of some languages allow for the subject of a tensed clause to remain unexpressed. The formulation of the Null Subject Parameter and the division between pro-drop and non-pro-drop languages are examined here. The second section provides an extensive overview of the formal/generative literature on the L2 acquisition of anaphoric subjects in Romance languages. The findings of early parametric studies are compared to those of the more recent 'syntax-discourse interface' literature. Finally, in the third section, some relevant discourse-oriented studies from variationist and pragmatic perspectives are reviewed.

3.1 Null subject and non-null-subject languages

In Spanish and Greek grammars, the explicit realization of subjects in finite clauses is not compulsory (for Spanish: Fernández Soriano, 1999; Luján, 1999; for Greek: Papadopoulou, Peristeri, Plemenou, Marinis, & Tsimpli, 2015; Tsimpli et al., 2004). Therefore, the grammatical subjects of Spanish and Greek sentences may remain unexpressed (null). In contrast, with the exception of a limited number of specific constructions (e.g. same-subject coordinate clauses) the same does not hold for English (Beavers & Sag, 2004; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Nariyama, 2004; Quirk, Greenbaum, Leech, & Svartvik, 1985). Consider the following examples (the imaginary context being about some person, e.g. María/Μαρία/Mary):

13) **ø** vino ayer. (Spanish)
ø ήρθε χτες. (Greek)
***ø** came yesterday. (English)
"She came yesterday"

Crucially, the null subjects in (13) are grammatically correct in the Spanish and Greek clauses, whereas the English clause is ungrammatical. Biberauer, Holberg, Roberts, & Sheehan (2010) provide an overview of the traditional accounts in formal linguistics (for Latin and Ancient Greek) regarding the grammaticality of null subjects in some languages. The authors cite the seminal study of Jespersen (1924) who noted that "in such languages (e.g. Latin) many sentences have no explicit indication of the subject" (p.213). However, this idea is much older. According to Biberauer and colleagues, since the times of the first syntactician in the history (Apollonius Dyscolus, 2nd century AD) it has been argued that "since a pronominal subject can be expressed 'in the verb' in languages such as Greek and Latin, there is no general requirement to pronounce the subject separately as a nominative pronoun" (Biberauer et al., 2010:3).

The traditional idea that in some languages with rich verbal morphology the subject may remain unexpressed was recovered by Perlmutter[40] (1971) and finally gave birth to the

[40] In formal/generative approaches there is a distinction between 'licensing' and 'identification'. In classic Generativism 'licensing' refers to the grammatical mechanisms that allow null subjects to occur whereas it is assumed that the 'identification' is made possible through verbal endings. It should be noted, however, that the 'subject in the verb' hypothesis cannot explain why some languages (such as Chinese) widely

'Null Subject Parameter' (NSP) under the generative linguistic framework[41] (Chomsky, 1981; Jaeggli, 1982; Jaeggli & Safir, 1989; Rizzi, 1986). Under this approach, languages are divided into two types: null-subject (or pro-drop) and non-null-subject (or non-pro-drop). Spanish, Greek and Italian are among the exemplary languages of the former type whereas English and French belong to the latter. Besides the possibility of null subjects, pro-drop languages are assumed to share a number of other characteristics, summarized by Montrul (2004:179) in the following table:

Setting	**+ pro-drop**	**- pro-drop**
language	Spanish	English
properties	rich verbal agreement inflection	poor verbal agreement inflection
	null and overt subjects	overt subjects
	null expletives	overt expletives
	preverbal and postverbal subjects	preverbal subjects
	that-t effect	**that*-t effect

Table 1. Properties of the pro-drop languages

permit null subjects whereas they lack a rich verbal morphology (Quesada, 2015:22). As Huang (2000:57) points out, "its prediction, that rich agreement systems constitute both a necessary and a sufficient condition for the licensing of referential null subjects, is empirically falsified in both directions". Whatever the case, it is out of the scope of this study to provide an explanation as to why null subjects are grammatically licenced in Spanish and Greek but not in English.

[41] The NSP constitutes the flagship of the 'Principles and Parameters' theory. According to Montrul (2004:178), it is "the first and most widely studied parameter in the theory". It should be noted that the original conceptualization of the parameter has undergone modifications under the Minimalist Program (Chomsky, 1995). However, it is out of the scope of this study to analyse in depth the corresponding generative accounts of the NSP. As Rothman (2009:953) points out: "Whatever analysis one takes is somewhat inconsequential for the immediate purposes inasmuch as it is observable that null-subjects are licensed in Spanish and that they co-exist with overt pronominal subjects".

Apart from these properties, pro-drop languages are also assumed to obey to the Overt Pronoun Constraint (OPC)[42]. Note here that Greek, in contrast with English, shares all the above pro-drop language properties with Spanish[43]. In addition to null subjects, the inventory of referential subject expressions of both pro-drop and non-pro-drop languages comprises pronouns (personal, demonstrative etc.) and lexical subjects (noun phrases, proper names etc.). Consider the following alternatives to example (13):

14) **Ella** vino ayer. (Spanish)
Αυτή ήρθε χτες. (Greek)
She came yesterday. (English)

15) **María** vino ayer. (Spanish)
Η Μαρία ήρθε χτες. (Greek)
Mary came yesterday. (English)

Whereas the repertoire of referential expressions is language-specific, the selection of one or another form in discourse is not arbitrary and strongly depends on the information status of the referent (see section 2.5). As we have seen, the more topical/given/active/accessible a referent (i.e. the higher its information status), the less informative the selected referential expression. Null subjects, as in (13), are commonly employed in Spanish and Greek whenever the speaker assumes the referent to be very prominent (e.g. in an answer to the question "When did she come?"). In this case, leaving aside stylistic and contrastive purposes, the overt pronoun in (14) and the proper name in (15), although perfectly grammatical, are overexplicit in the Spanish and Greek

[42] The OPC accounts for the fact that, in pro-drop languages, overt pronouns cannot be linked to a quantified antecedent such as in this example: *Nadie$_i$ dice que* ***el**$_{*i}$ lo sepa todo* ("Nobody$_i$ says that **he**$_{*i}$ knows everything"). It is out of the scope of this study to provide an extensive description of the OPC given that we focus only on the first property of null-subject languages, namely the alternation of referential null and overt subjects. For more information on the original formulation of the OPC see Montalbetti (1984). For the acquisition of the OPC in L2 Spanish see Lozano (2002b, 2002c, 2008).

[43] Note, however, that even between null-subject languages, some micro-variation in the properties of referential subjects may exist. Recently, Papadopoulou et al. (2015) found cross-linguistic differences in the interpretation of overt subject pronouns in Greek and Spanish. Additionally, Carminati (2002) and Filiaci (2010) also provided evidence regarding similar differences between Spanish and Italian. Despite the differences, however, it is the widespread omission of subjects that sharply characterizes pro-drop languages (Spanish, Italian and Greek) and separates them from English and other non-pro-drop languages.

examples[44]. In contrast, the corresponding referential choice for a salient referent in English is the overt pronoun, given the lack of other less informative expressions (since null subjects are ungrammatical in English).

Regarding anaphoric subjects, the different instantiation of the NSP creates a striking asymmetry between pro-drop languages (such as Spanish and Greek) and non-pro-drop languages (such as English). This asymmetry is directly reflected in discourse. Whereas overt pronouns constitute the prototypical anaphoric subject expressions in English, the corresponding more common referential choice in pro-drop languages is the null subject. More than half of the 3rd person anaphoric subjects (60%) in the native Spanish data of this study were found to be null (see section 6.1). A proportion range of unexpressed subjects in Spanish discourse of about 60%-80% is reported in several other corpus studies[45] (Geeslin & Gudmestad, 2008; Gudmestad & Geeslin, 2010; Montrul & Rodríguez Louro, 2006; Quesada & Blackwell, 2009; Bel et al., 2010). This high proportion of 3rd person null anaphoric subjects is also observed in Greek L1 discourse. Charatzidis, Georgopoulos, Papadopoulou, & Tantos (2015) examined Greek L1 texts and found that 67% of the 3rd person subjects remain unexpressed. Note that the proportion of overt pronouns in the Greek L1 data of the aforementioned study is only 4%. This is not surprising given that in null subject languages, overt pronouns are rarely used (Arnold, 1998:92; Bel & García-Alcaraz, 2015:228; Shin & Cairns, 2012:30). In the results of the present thesis, regarding the Spanish L1 data, only 7% of the anaphoric subjects were expressed pronominally. Similar proportions (around 10%) are reported in other corpus studies as well (Geeslin & Gudmestad, 2008; Gudmestad & Geeslin, 2010). At the same time, the corresponding pronominal rates in English L1 data has been found to be above 90% (Shin & Montes-Alcalá, 2014; Torres Cacoullos & Travis, 2014). Given this obvious asymmetry between pro-drop and non-pro-drop languages, it is reasonable to assume that English-speaking learners of L2 Spanish may experience some difficulties in the production of anaphoric subject expressions due to negative cross-linguistic influence. In particular, they may produce more overt pronouns than necessary (which

[44] The term 'overt pronoun' is employed in the formal literature under the assumption that the unexpressed subjects are also pronouns ('null pronouns').

[45] Some variation in the rates of null subjects should be expected due to the different discourse genres examined in each study (Travis, 2007).

may lead to overexplicitness, as defined in 2.4.3). One of the aims of this thesis is to explore whether, and to what extent, they do so. Additionally, their data are compared to those of Greek-speaking learners of the same proficiency level in order to examine whether the Spanish/Greek similarity with respect to anaphoric subjects may be a facilitating factor for the Greek-speaking groups. Before we proceed, a detailed review of previous anaphora studies in SLA is in order and will be carried out in the following sections.

3.2 Formal/generative SLA studies on anaphora

In the last three decades, based on the original formulation of the NSP under the generative framework, several authors set forth to examine the potential acquisitional problems related to the interpretation and distribution of anaphoric subject expressions. The first SLA studies were conducted under the 'Principles and Parameters' framework (Chomsky 1980, 1981) and exploited the NSP to investigate the possibility of access to Universal Grammar (UG) in the L2. The acquisition of anaphoric subjects was examined in purely syntactic terms in the early parametric SLA studies, in sharp contrast to other coevally conducted discourse-oriented studies on anaphora in L1[46] (Chafe, 1980; Givón, 1983 *inter alia*). In the next decade, the original formulation of the NSP was further developed under the Minimalist Program (Chomsky, 1995) and very recently, drawing heavily on modular views of language (Jackendoff, 1997), the focus of interest in generative SLA approaches has shifted towards the linguistic interfaces (for overviews see: Montrul, 2011; Rothman & Slabakova, 2011; White, 2009). Anaphoric subjects have been extensively studied in the last few years due to their prototypically 'interface' nature, namely the fact that their use (interpretation and distribution) is assumed to be constrained by both syntax and discourse (i.e. they "fall squarely" under the syntax-discourse interface, as Montrul (2011:603) points out). In the following sections, an extensive review of the formal/generative studies on the acquisition of anaphoric subjects will be performed.

[46] Recall that the theoretical discourse-oriented studies on anaphora focused on native speakers and were not concerned with acquisitional issues (see Chapter 2).

3.2.1 Early studies on the acquisition of the NSP

The first generative studies on the acquisition of the NSP were mainly concerned with three issues. First, they attempted to determine which option of the parameter (whether the pro-drop or the non-pro-drop) is the default one ('unmarked') in SLA. Second, they aimed to test the assumed clustering of properties of the parameter. Finally, they sought to understand whether L2 learners are able to reset the parameter. Although none of the above issues is addressed in the present study (at least not in these terms), the examination of the early approaches is necessary since they provide important evidence with respect to the acquisition of the NSP. In these studies, the role of transfer is examined alongside with the eventual access L2 learners may have to UG. It should be noted that, due the binary definition of the parameter, null subjects are examined only in opposition to overt pronominals. The rest of alternative anaphoric forms (notably noun phrases) are largely overlooked in the early studies on the acquisition of NSP in L2 Spanish. Additionally, early studies focused exclusively on the grammaticality of the anaphors (and not to their pragmatic felicity). The role of discourse/pragmatics was occasionally mentioned but, crucially, never empirically tested.

Interestingly, one of the first formal studies in the acquisition of the pro-drop parameter in Spanish L2 underlines the fact that referential null subjects are pragmatically constrained (Liceras, 1988). The author argues for the separate analysis of the clustered properties that have been assigned to the null-subject parameter, given that some of them are regulated by syntax (e.g. obligatory expletives) whereas others obey to stylistic purposes (e.g. 3rd person referential subjects). Liceras highlights the fact that "a subject pronoun can be optionally or obligatory deleted depending on the type of construction" (p.75). According to the author, a null pronoun is obligatory in Spanish with existential or weather verbs such as in the following example (Liceras, 1988:81):

16) *__ELLO__ hace un viento horrible
"__It__ is terribly windy"

In contrast, the overt pronoun in example (17) is not obligatory (in fact, according to the author, the overt pronoun in this case is redundant):

17) Pedro$_i$ está muy cansado. **#Él$_i$** ha dormido como un cesto.
"Pedro$_i$ is very tired. **#He$_i$** has slept like a log."

The author examined two French and two English advanced learners of Spanish by means of a story-telling task and a grammaticality judgment test (GJT). Results indicated that

learners have acquired the 'pleonastic pro' (obligatory null expletives) and the 'pro-drop' (optional null subjects) features[47]. High production of redundant subjects, however, was detected in the stories of one of the four participants. This was attributed to the nature of the specific story. In a following-up study (Liceras, 1989) the author focused exclusively on the acceptability of ungrammatical overt pronouns in contexts where null subjects are obligatory in Spanish (null expletives in existential and weather verbs such as in example (16) above). Liceras examined learners with French and English L1 background at several levels of proficiency in Spanish L2. The fact that the learners rejected ungrammatical expletive pronouns was interpreted as evidence in favour of the resetting of the NSP. According to the author, no transfer effects should be expected with null subjects since "pleonastic *pro* is incorporated in the learners' grammar at the very early stages" (Liceras, 1989:126). Note however that discursively constrained anaphors (e.g. 3rd person referential subjects) were not examined in this study. Liceras's findings may be contrasted to White (1985, 1986) who, in the opposite direction, tested Spanish, Italian and French learners of English L2 at several proficiency levels (from beginner to advanced). She found that the Spanish participants, unlike their French colleagues, in many cases failed to detect the ungrammaticality of sentences where the subject was missing (obligatory null expletives such as in example (16) above). This was interpreted as an indication of negative cross-linguistic influence. In her own terms, "the results of this study indicate that having to change a parameter of UG causes problems for language learners and that this is a source of transfer errors, particularly at lower levels of proficiency" (White 1985:60). As in the case of Liceras (1988, 1989), however, White examined the role of transfer in anaphora resolution only in terms of the acceptability of ungrammatical sentences.

One of the first production-oriented studies in Spanish L2 anaphora was carried out by Phinney (1987). In line with Liceras (1988, 1989) and White (1985, 1986), the author

[47] Within the Principles and Parameters framework, it is broadly assumed that the mere acceptance/production of some null subjects constitutes evidence that the pro-drop parameter has been reset. However, as Saunders (1999:13) points out, "unless learners understand the disambiguating and emphatic functions of overt pronouns in Spanish, as well as the distinctive functions of pronouns and noun phrases, it cannot be claimed that null subjects have been acquired".

aimed to apply the parameterized model to SLA by testing the role of the NSP in the acquisition of English and Spanish. Phinney examined the written production (compositions written in 1st person) of beginners and low-intermediate learners in both directions (Spanish L1/English L2 and English L1/Spanish L2) and concluded that English-speaking learners of L2 Spanish seem to reset the parameter without problems, since they consistently omitted 1st person subjects. Spanish-speaking learners of L2 English, on the other hand, have long-lasting difficulties with subject pronoun usage, since they produced some ungrammatical 1st person null subjects in English [48]. Additionally, Phinney provided an interesting overview of the early Contrastive Analysis (CA) and Error Analysis (EA) approaches in SLA. She cited Stockwell, Bowen, & Martin (1965:421) who, under the CA approach, proposed that English-speaking learners of Spanish "will tend to overuse subjects – and in doing so will sound emphatic and aggressive". More specifically, 3rd person subject omission in Spanish was considered by the authors to be the second more important learning problem for English L1 speakers (after phonological problems). In contrast, 1st and 2nd person subject omission was considered to be far easier[49] (position 14 in a 16-point hierarchy of difficulty). However, after examining more recent studies and in line with her results, Phinney concluded that 3rd person subject omission does not pose a sufficiently serious problem for the English speakers since "with regard to the overuse of subjects hypothesized by Stockwell *et al* (1965), there appears to be a conspiracy of silence. None of the studies cited mentioned overuse of pronominal subjects as an error" (p.232). Apart from the fact that Phinney examined only 1st person pronouns in her study (which were not considered especially problematic by Stockwell and colleagues), the key to the interpretation of her data lies in her last observation. Crucially, the overuse of pronominal subjects in Spanish is not a grammatical error (leaving aside obligatory null expletives). When examined in grammaticality terms, all referential subject expressions are correct and the fact that only discourse/pragmatic factors may account for their felicity was overlooked in Phinney's study.

[48] It should be noted that no examples from the L2 production data are provided by the author for any of the examined language combinations.

[49] Recently, this claim was empirically demonstrated by Lozano (2009b).

Another particularly relevant study in the production of anaphoric subjects in SLA contexts is Bini (1993). In line with the above-mentioned studies, the author examined the acquisition of the NSP by Spanish-speaking learners of Italian (beginner and low-intermediate levels of proficiency). Crucially, both languages are pro-drop and are assumed to share the same characteristics of the NSP. Bini (1993:128) argued for the need to examine subject pronouns in a real communicative situation given that "en italiano como en español la presencia de los pronombres sujeto depende de factores estilísticos y pragmáticos" [50]. Therefore, she tested the production of 1st and 2nd person subject expressions in interview data and found that the Spanish-speaking learners (mostly the beginners) overuse both pronouns ('io' and 'tu') in Italian L2. Given that the role of negative grammatical transfer is discarded in this case, the author provided a discourse/pragmatic interpretation for her findings. First, assuming that oral discourse production is difficult for L2 learners, Bini argues that these may use the redundant pronouns in order to have more time to think. Second, assuming that Spanish students are conscious of their limited competence in Italian, they may prefer the more explicit version of the subject expression in order to avoid potential ambiguity (due to their incomplete knowledge of verb morphology). Bini (1993), in line with Liceras (1988), highlighted the need of considering discourse/pragmatic factors in the acquisition of the NSP. Her findings regarding lack of positive transfer effects might be interpreted as running against White (1985, 1986) who found that French-speaking learners of English perform better than their Spanish-speaking counterparts. It should be noted, however, that Bini did not include a comparable L2 group with non-pro-drop L1 background in her study and that the detected overexplicit production concerns only low proficiency groups. Additionally, what Bini actually demonstrated is lack of categorical positive transfer effects in intermediate learners when both source and target languages are pro-drop. This finding is not directly comparable to the negative transfer effects when the L1 and L2 differ, as reported in White (1985, 1986). More comparable to the latter study is Polio (1995), who tested the oral production (film retellings) of English (non-pro-drop) and Japanese (pro-drop) learners of Chinese (pro-drop). The author found that, regarding the production of null subjects, the learner groups did not differ with each other. This finding runs against the transfer-related claims in White (1985, 1986). Polio also found that both learner

[50] "In Italian, like in Spanish, the presence of subject pronouns depends on stylistic and pragmatic factors" (translation mine).

groups overuse pronouns and noun phrases with respect to the native speakers. The author interpreted this finding in the same line with the explanations of Bini (1993). Additionally, she considered the role of classroom input as a potential factor: "the input to the students may contain more pronouns than found in everyday speech to NS" (Polio, 1995:372). Finally, the author hypothesized that the increased use of pronouns in the production of the Japanese learners may be due to the fact that all of them had studied at least some English (L3 transfer).

Fernandez de Moya (1996) was the first, to my knowledge, to examine Spanish L2 learners with Greek L1 background. The author compared Spanish L2 interview data of monolingual learners with different L1s (English, French) and bilinguals (English/Arabic, English/Portuguese and English/Greek) in order to examine the role of transfer in the acquisition of verb morphology and the NSP. The qualitative analysis of the data (students were examined one by one) revealed some important trends. Regarding the English and French learners, the author observed variability in the production of null subjects, insofar as some students seem to rely on the properties of their L1 (overproduction of overt pronominals) whereas others do not. Crucially, the lowest proportion of null subjects in the study was found in the data of a monolingual English-speaking learner. On the other hand, the overall production of bilingual learners with Arabic, Portuguese and Greek (all of them pro-drop) backgrounds revealed some potential cross-linguistic influence. Crucially, the English/Greek bilingual learner was the one with the highest production of null subjects. Specifically, she/he was the only participant that did not employ any overt pronouns at all in the interviews. Additionally, her/his production of verb morphology errors was the lowest among all the participants of the study (the highest being that of a monolingual English speaker). The author concluded, in line with Bini (1993), that the incomplete knowledge of verb morphology goes hand in hand with the overproduction of overt pronominal subjects. Additionally, the role of the L1 was considered crucial as a conflating factor which may facilitate or hinder the process of the acquisition of both the aforementioned linguistic phenomena.

Similarly to Fernandez de Moya (1996), but from a different theoretical perspective, Liceras & Díaz (1999) also examined L2 Spanish learners from several L1 backgrounds. More specifically, the authors tested the oral production of English, French, German, Chinese and Japanese upper-intermediate learners of L2 Spanish. In the overall distribution of subject forms, no significant differences were found between the learners and the native control group, i.e. all L2 groups were found to produce native-like

proportions of null subjects (both in main and subordinate clauses). More specifically, learners from a non-pro-drop L1 background (English L1 and French L1) did not overproduce overt pronominal subjects (with the exception of one French-speaking learner). Similarly, learners from a topic-drop L1 background (German L1, Chinese L1 and Japanese L1) were found to produce null subjects both in main and in subordinate clauses. The authors argued that their findings regarding the production of null and overt subject pronouns are "best explained by assuming that the adult IL grammar has a default licensing procedure which is responsible for the production of null subjects provided they are identified" (Liceras & Díaz, 1999:35).

In a more recent study, Lozano (2002a) corroborated that low proficiency English-speaking learners of Spanish reset the NSP by examining the acceptability rates of ungrammatical overt expletives (ExpS) and referential null/overt subjects (ProS) in paired grammaticality judgment tests (GJTs). The author found that, although learners seem to initially transfer the parametric setting from their L1, they gradually behave more native-like by rejecting ExpS (such as the ones in example (16) above). Regarding the alternation of ProS, see the following example from Lozano (2002a:48):

18) **Yo**/**Ø** voy a la universidad en coche.
"**I** go to the university by car."

Learners were found to accept both null and overt pronouns, such as the ones in example (18), from the first stages of acquisition. The gradual resetting of expletives against the instantaneous resetting of the pronominal subjects was taken as evidence that learners may have different representation of each property of the NSP. Additionally, Lozano was one of the first to point out the need for establishing criteria regarding the distribution of null and overt subjects "so that L2 learners can be taught precisely under what conditions ProS can be used in Spanish" (p.55).

3.2.2 The acquisition of the NSP and the role of discourse[51]

The importance of discourse, largely overlooked during the first two decades of investigation on the acquisition of anaphoric subjects in SLA, gradually begun to be addressed in more recent studies. With respect to this, White (1989:86) had made early

[51] The terms 'discourse' and 'pragmatics' are used interchangeably here, following the predominant tendency in the SLA literature (White, 2011:581).

on the following crucial observation: "there are two things that an L2 learner of a [+pro-drop] language has to acquire: (i) the fact that null subjects are permitted, and (ii) the circumstances in which the language actually makes use of the fact that null subjects are permitted". Recall that, in early SLA studies on the acquisition of the NSP, these circumstances were not empirically examined. This point is also made explicit in Williams (1988:342) who argues that "because the aim of these UG studies has generally been to examine linguistic competence, discourse function plays little role, and no explanation is provided for the distribution of omitted pronouns". Additionally, Polio (1995:354) points out that "simply presenting the number of zero subject pronouns used by NNs tells us little without a native speaker comparison".

In more recent formal studies in the acquisition of the NSP, the role of discourse was partially upgraded. Al-Kasey & Pérez-Leroux (1998) focused on the interpretation and production of expletives (obligatory null) and referential subjects (optionally null) by English-speaking learners of Spanish at five proficiency levels. It should be noted here that this is one of the first NSP acquisition studies that includes a native control group in the methodological design. The authors recognized the role of discourse for the referential null subjects: "subjects may be omitted only when they are recoverable from the context" (p.164). However, in their study it is not determined (and consequently not tested) under what contextual conditions the subjects may be omitted. Instead, in the interpretation task, the authors compare the acceptability of ungrammatical expletive pronouns (such as the ones in example (16) above) by the learners and the native control group. Regarding the production task, the focus of interest was again on the grammaticality of the subject expressions. The authors concluded that the NSP can be reset in the case of both null expletives and referential subjects, since no grammatical deficits were found for the advanced English-speaking learners. These results are in line with the findings of Liceras (1988, 1989) and Phinney (1987).

In line with Al-Kasey & Pérez-Leroux (1998), Pérez-Leroux & Glass (1997, 1999) looked at the acquisition of NSP properties by English-speaking learners of Spanish. In the first of the two studies (Pérez-Leroux & Glass, 1997), focusing exclusively on highly proficient learners in opposition to a native control group, the participants were asked to complete contextualized sentences with null and overt subjects in two different tasks. First, an OPC story, where the null subject choice is assumed to be syntactically bound (see section 3.1 for more details on the OPC). Second, a topic/focus story, where the

alternation of null/overt subjects is assumed to depend on the traditional old/new information distinction, as in the example below[52] (Pérez-Leroux & Glass, 1999:236):

19) **Topic/focus story**
Hace calor y la familia va al jardín.
"It is hot and the family goes out to the garden."
Subject question:
¿Quién piensa la abuela que regará las plantas?
"Who does the grandmother think will water the plants?"
Target focus response:
La abuela piensa que **ella** regará las plantas.
"The grandmother thinks that **SHE** will water the plants"
(embedded subject is focused)
Object question:
¿Qué piensa la abuela que hará en el jardín?
"What does the grandmother think that she will do in the garden?"
Target topic response:
La abuela piensa que **ø** regará las plantas.
"The grandmother thinks that (**she**) will water the plants.'
(embedded subject is topic)

The learners were found to produce more null subjects than the native speakers in both tasks. According to the authors, this suggests that they master the discourse/pragmatic properties related to the distribution of null subjects in L2 Spanish. In a follow-up study (Pérez-Leroux & Glass, 1999) the authors examined English-speaking learners of Spanish at three proficiency levels in comparison with a native control group. They tested the syntactic violation of the OPC in the translation of sentences from English to Spanish and the discourse-constrained production of null subjects in topic/focus stories (such as the ones in example (19) above). Regarding the first task, the authors found no significant differences between learners and native speakers. Regarding the second task, the authors found that the performance of the native control was significantly different from the

[52] The authors claim that, in pro-drop languages such as Spanish, "syntactically, null pronouns are barred from certain positions" and that "a null subject is not possible when the information provided by the subject is taken as new within discourse" (Pérez-Leroux & Glass, 1999:226). The former claim is not true regarding the phenomenon under study, since null subjects are always syntactically allowed in pro-drop languages. Regarding the latter claim, as it has been discussed, the definition of topic/focus in terms of old/new information may pose methodological problems since it does not say much about the distribution of null subjects in real discourse (see section 2.4.1).

performance of the elementary and intermediate L2 groups, but not from the performance of the advanced L2 group. More specifically, elementary and intermediate learners were not native-like in their selection between null and overt subjects for the topic and focus stories, respectively. In line with Al-Kasey & Pérez-Leroux (1998), they concluded that the overall use of null subjects increases with language experience and that learners have more difficulty mastering their distribution at the discourse level. The authors made additionally two crucial observations. In line with Lozano (2002a), they noted that "the most relevant data is not null pronoun use *per se* but the level of discrimination between semantic contexts in pronoun use" (p.242). Additionally, they acknowledged that transfer could not be considered in their study, due to the L1/L2 mismatch in the NSP properties. In line with this, it may be argued that, in the traditionally studied English L1/Spanish L2 combination, the role of potential cross-linguistic influence may never be completely discarded.

In a similar fashion with Pérez-Leroux & Glass (1997, 1999), Lozano (2002b, 2002c) examined the acquisition of properties related to the NSP in Spanish L2. Importantly, Lozano's studies are among the first to follow an infrequently employed methodological approach where L2 learners from one pro-drop and one non-pro-drop L1 background are directly contrasted (Polio, 1995; White, 1985, 1986). The author used an acceptability judgment task (AJT) to examine advanced Greek-speaking and English-speaking learners of Spanish in two different constructions in Spanish[53]. Apart from the well-studied syntactically-bound OPC construction (Pérez-Leroux & Glass, 1997, 1999), Lozano examined patterns regulated by the contrastive focus constraint (CFC) where an overt pronoun is required for the felicitous interpretation of the sentence. Crucially, this constrain is operative in both Spanish and Greek, but not in English, as we see in the example that follows (Lozano, 2002c:55):

20) Context: Mr López$_j$ and Ms García$_k$ work at the university and at a famous publishers. However...

a. cada estudiante dice que **él$_j$/#pro$_j$** tiene poco dinero.(Spanish)

b. o kathe mathitis lei pos **aftos$_j$/#pro$_j$** ehi liga lefta.(Greek)

[53] It should be noted that this is one of the first SLA studies that acknowledges the methodological limitations related to the differences between anaphoric interpretation and production. In terms of the author: "Only interpretation tasks were used, which says nothing of the learners' production of pronominal subjects" (Lozano, 2002b:64).

c. each student says that **he$_j$/*pro$_j$** has little money.(English)

The author found that both learner groups behave native-like in the OPC contexts. This is in line with the findings of Pérez-Leroux & Glass (1997, 1999). On the other hand, in the CFC contexts, the English group was found to significantly differ from the native speakers in the acceptability of null pronouns (some English speakers accepted null pronouns in CFC contexts). Crucially, the Greek-speaking learners behaved like native speakers in the acceptability of the same constructions. Based on these results, the author argued that both UG and transfer may account for the acquisition of the NSP properties to conclude that "L1 is the key to representational deficits at the advanced levels of proficiency" (Lozano 2002b:65). This might be considered to run against the findings of Bini (1993) but, crucially, the learner participants of the two studies differ with respect to proficiency. The results of Lozano, however, are in line with White (1985, 1986) who also found negative cross-linguistic influence to constrain anaphoric choices of L2 learners. Future studies on the OPC (Lozano, 2008b; Rothman & Iverson, 2007a, 2007b, 2007c) clearly demonstrated, in line with the earlier studies (Lozano, 2002b, 2002c; Pérez-Leroux & Glass, 1997, 1999), that English learners of Spanish show native-like knowledge of this syntactic constraint from early developmental stages. More recently, regarding the CFC, Rothman (2007, 2009) tested English-speaking learners of Spanish and found, in line with Lozano (2002b, 2002c), that the intermediate proficiency groups accept more contextually infelicitous null subjects than the native Spanish speakers.

3.2.3 Summary of results of the early formal literature

Leaving aside concerns that are out of the scope of this thesis (regarding the markedness and the clustering of properties of the NSP), the summary of first two decades of investigation on the acquisition of the NSP highlights two important issues. On the one hand, that the role of cross-linguistic influence remains unresolved. Some studies claim to have demonstrated the resetting of the parameter (Al-Kasey & Pérez-Leroux, 1998; Liceras, 1988, 1989; Lozano, 2002a; Pérez-Leroux & Glass, 1997; Phinney, 1987) based on evidence indicative of the awareness of learners that the subject position of a tensed clause in pro-drop languages may remain empty. Other studies resort to the traditional hypothesis of potential cross-linguistic influence (Fernandez de Moya, 1996; Lozano, 2002b, 2002c; White, 1985, 1986) in order to explain the acceptance/production of ungrammatical null subjects by learners. Crucially, all the studies reviewed so far have focused mostly on the syntactic properties of the anaphoric subjects. On the other hand,

there is an increasing awareness in the early literature regarding the crucial role of discourse (Bini, 1993; Liceras, 1988; Lozano, 2002a, 2002c; Pérez-Leroux & Glass, 1999) which will be more successfully addressed in the more recent syntax-discourse interface studies.

3.2.4 More recent formal approaches: the syntax-discourse interface

The focus of interest in the acquisition of anaphoric subjects has drastically changed in the last two decades or so. Whereas the importance of discourse in anaphora was barely hinted in the early formal studies, more recent SLA research has widely acknowledged the fact that the interpretation and distribution of null and overt anaphoric subjects is highly dependent on contextual information[54]. As we have seen (section 2.3) the present study is in line with this point of view. This approach coincides with some influential accounts in theoretical linguistics which propose the conceptualization of the language faculty in terms of the interaction between modules (Jackendoff, 1997, 2002; Ramchand & Reiss, 2007). According to these accounts, the different components of language (syntax, semantics, phonology) interact with each other as well as with other aspects of general cognition (discourse/pragmatics) and this (point of) interaction is accounted for in terms of 'interfaces'[55]. In SLA, several linguistic phenomena related to the interfaces have been extensively studied under the assumption that their acquisition may be particularly problematic for L2 learners (in terms of 'optionality and 'variability')[56]. The bulk of the relevant literature focuses on the particular interface between syntax and

---

[54] Notice that outside generative approaches, anaphora is widely assumed to be highly constrained by discourse, in terms of information structure (see Chapter 2). This is the exact opposite view of the one held in the early parametric studies on the acquisition of the NSP where the interpretation/production of anaphoric subjects was treated in grammaticality terms.

[55] The term 'interface' may refer to both (a) the link between the different linguistic modules (internal interfaces) and (b) the link between linguistic modules and general cognition (external interfaces) (Ramchand & Reiss, 2007).

[56] 'Optionality' or 'variability related to the interfaces has also been reported for L1 attrition (Montrul, 2004a; Sorace, 2004; Tsimpli, Sorace, Heycock, & Filiaci, 2004) and bilingual acquisition (Bel, García-Alcaraz, & Rosado, 2016; Paradis & Navarro, 2003; Pladevall-Ballester, 2009; Serratrice, 2007a; Serratrice, Sorace, & Paoli, 2004; Sorace & Serratrice, 2009; Sorace, Serratrice, Filiaci, & Baldo, 2009). Due to the specific focus of this thesis, however, no particular emphasis shall be given here to domains outside adult L2 acquisition.

discourse (White, 2011:580), whereas the interpretation and distribution of anaphoric subjects is the most widely examined linguistic phenomenon in the syntax-discourse interface literature in SLA (Montrul, 2011:594).

The Interface Hypothesis (IH), originally proposed by Sorace and colleagues (Sorace, 2004, 2005, 2006a; Sorace & Filiaci, 2006), aims to account for the non-target behaviour of highly advanced L2 learners regarding specific linguistic phenomena (notably AR). According to the author (Sorace, 2011:5), the original strong version of the IH "predicts that structures involving an interface between syntax and other cognitive domains present residual optionality (in L2 acquisition)". Regarding the prototypical syntax-discourse phenomenon of anaphoric resolution/distribution, the direct implication of the IH is that L2 learners are expected to behave differently from native speakers even at the highest levels of proficiency. More specifically, the IH predicts that even the near-native L2 learners may never be native-like with respect to the interpretation and production of anaphoric subjects (Sorace, 2011:26; White, 2016:29). The IH led to the formulation of several claims regarding the sources of this residual optionality. Most importantly, whereas the original version does not exclude the possibility of cross-linguistic influence (Sorace & Filiaci, 2006:343), later versions of the IH assume the syntax-discourse interface to be unaffected by transfer[57] (Sorace & Serratrice, 2009:207). In other words, highly advanced L2 learners are expected to exhibit non-target performance with syntax-discourse interface phenomena such as anaphora, irrespectively of their L1. At the same time that the role of transfer is considerably (or completely) diminished, residual difficulties of L2 learners are attributed to the processing cost of accessing and integrating information from the two modules (syntax and discourse) involved in the interface[58]. Very recently, Sorace (2016) highlighted the relevance of two factors involved in the processing difficulties at the syntax-discourse interface, namely: competition of resources

[57] Recently, Sorace (2011, 2012) acknowledged that several cumulative factors (including transfer) may simultaneously be at play in the acquisition of linguistic phenomena at the interfaces. In other words, the claim of 'no transfer' was converted to 'not *only* transfer' in the latest versions of the IH (in line with the original version).

[58] The 'processing resources' account of the IH is partially in line with the 'Shallow Processing' hypothesis (Clahsen & Felser, 2006). Crucially, however, in the latter it is assumed that the increased processing cost is the result of target-deviant grammatical representations whereas, in the former, defective processing originates from the endeavour related to the integration of multiple modules.

and cognitive load. According to the author, L2 learners are among the populations that are especially sensitive to these factors due to the need of applying a considerable amount of inhibitory control to their dominant L1 when they use the L2. As a result, the author argues that "the overt pronoun may be a default form used to relieve processing demands when these become temporarily unmanageable" (Sorace, 2016:676). It should be noted, finally, that processing-related difficulties are expected to be manifest both in online and offline data, as Sorace (2011:20) points out: "It is important to dispel a frequent misconception that only online tasks can test the processing resources account. In fact, both offline and online tasks give insights about speakers' processing abilities". All in all, as we will see in the literature review that follows, the bulk of evidence supporting the IH comes from offline data.

In contrast to the processing explanation, the role of cross-linguistic influence is highlighted in two influential representational accounts which shall be examined here. On the one hand, Tsimpli et al. (2004) proposed the Underspecification Account (UA), according to which the target-deviant behaviour of L2 learners results from the underspecification of interpretable features[59]. Regarding the acquisition of anaphoric subjects, for example, English-speaking learners of a pro-drop L2 such as Italian and Spanish, may erroneously map the interpretable feature [+Topic Shift] ([+TS]) to both null and overt subjects, whereas the same feature is mapped only to overt pronouns in the target language[60]. This is due to cross-linguistic influence from their L1, where overt pronominal subjects (in the absence of nulls) are used indistinctively (i.e. they are not 'specified') for [+TS] and [-TS]. The result is the overacceptance/overproduction of infelicitous overt pronouns. The limitation of this approach, as noted by Sorace (2011:13), is that it fails to account for the overextension of overt pronouns by learners whose L1 has the same mapping of interpretable features with the L2 (e.g. Greek L1/Spanish L2). On the other hand, the Interpretability Hypothesis (IPH), originally proposed by Tsimpli & Dimitrakopoulou (2007), is an alternative representational account according to which

[59] In recent theoretical accounts (*inter alia* Chomsky, 1995), linguistic features are divided into two types: interpretable features, which are related to meaning/semantics and uninterpretable features, which are purely syntactic/grammatical (Leal Mendez & Slabakova, 2014:538).

[60] It should be noted that the one-to-one mapping assumed by the UA (null pronouns for [-TS] and overt pronouns for [+TS]) is subject to the discussed limitations regarding the problematic binary conception of topic (see section 2.4.1).

the attested problematic L2 performance is due to the incomplete acquisition of formal uninterpretable features. More specifically, regarding the acquisition of anaphoric subjects, the non-target behaviour of L2 learners is attributed to the impossibility of resetting the NSP parameter when the L1/L2 pairing is syntactically different (Tsimpli & Roussou, 1991). In other words, the role of syntax is highlighted in the IPH, insofar as the parametric options of the L1 are assumed to decisively constrain L2 representations, leading to persistent problems. Notice, however that the bulk of the evidence supporting the IPH is based on the acceptance/production of ungrammatical null subjects in non-pro-drop L2 (such as English) by learners with pro-drop L1 (such as Spanish and Greek). As with the UA, this approach may not account for potential deficits when the L1/L2 combination shares the same parametric options (e.g. Greek L1/Spanish L2).

In sum, a considerable body of SLA investigation has recently focused on the well-attested observation that L2 learners show persistent deficits ('fossilization' as coined by Selinker, 1972) regarding the acquisition of particular linguistic features at the syntax-discourse interface. The anaphoric resolution/distribution of subject expressions is a phenomenon that has received an important part of this attention. Currently, there are two lines of explanations regarding the non-native performance of L2 learners in the use and interpretation of anaphoric subject expressions. On one side, the IH is a processing account which claims that, given the nature of the phenomenon, the cost of integrating information from two different modules (syntax and discourse) may saturate the computational system and lead to the attested optionality in L2 performance. On the other side, the representational accounts point to the role of cross-linguistic influence in order to provide explanations for this problematic behaviour. Two sorts of claims have been made with respect to this. First, the UA suggests that the root of the problem lies in the erroneous mapping of uninterpretable discourse-related features when the language pairing is asymmetrical (non-pro-drop L1/pro-drop L2). Second, similarly to the role given by the UA to the role of transfer, the IPH proposes that the variability is due to interpretable features such as the original setting of the NSP. The literature review presented in the following section aims to shed light on the evidence that supports one or another of the aforementioned accounts.

3.2.5 L2 acquisition studies on anaphora at the syntax-discourse interface

The IH was originally proposed as such in an L2 acquisition study by Sorace & Filiaci (2006). The authors examined the interpretation of globally ambiguous null and overt

pronouns in forward and backward anaphora by near-native speakers of Italian with English L1[61]. More specifically, based on the same experimental design of Tsimpli et al. (2004), they used a Picture Verification Task (PVT) in order to test the AR preferences of learners and native speakers of Italian. The authors aimed to test the Position of Antecedent Strategy (PAS) in Italian L2, initially proposed by Carminati (2002) to account for 3rd person anaphoric subjects in Italian L1. According to the PAS hypothesis, null pronouns are preferably interpreted as coreferential with an antecedent in Spec IP (the subject of the previous/next clause) whereas overt pronouns prefer a non-subject antecedent assignment[62], as in the examples from forward and backward anaphora that follow (Sorace & Filiaci, 2006:352):

21) (forward anaphora)

La mamma$_i$ dà un bacio alla figlia$_k$ mentre **lei$_{k/l}$/pro$_i$** si mette il cappotto.

"The mother$_i$ kisses her daughter$_k$, while **she$_{k/l}$/pro$_i$** is wearing her coat."

22) (backward anaphora)

Mentre **lei$_{k/l}$/pro$_i$** si mette il cappotto, la mamma$_i$ dà un bacio alla figlia$_k$.

"While **she$_{k/l}$/pro$_i$** is wearing her coat, the mother$_i$ kisses her daughter$_k$."

The results of Sorace and Filiaci (2006) indicate that the interpretation preferences of learners and native speakers are not identical in all cases. Regarding the null subjects, on one side, the English-speaking learners displayed similar preferences to the Italian natives in both forward and backward anaphora sentences. The learners, however, behaved differently from the natives in the case of the overt anaphors. The former group interpreted significantly more overt pronouns as being coreferential with the subject

[61] Forward anaphora in Sorace & Filiaci (2006) refers to cases where the clause with the antecedent precedes the clause containing the anaphor (antecedent before anaphor): e.g. "John$_i$ listens to music while **he$_i$** drives". Backward anaphora, according to the authors, refers to the reverse order (anaphor before antecedent): e.g. "While **he$_i$** drives, John$_i$ listens to music". However, there is no consistent definition of the two types of anaphora in the literature. Tsimpli et al. (2004), for example, employ the terms with the exactly reverse meaning.

[62] Notice that the PAS may account for the interpretation of 3rd person ambiguous anaphors in isolated experimental sentences consisting of two clauses where exactly two same-gender referents are involved. In real discourse, however, anaphoric patterns are often more (or less) complex than that and several other factors (apart from the structural position of the antecedent) are at play (see section 2.5).

antecedent than the latter. This target-deviant behaviour, despite the near-native proficiency level of the learners, was accounted for by the authors in terms of persisting indeterminacy at the syntax-discourse interface due to the lack of processing resources (in line with the IH). These results are also consistent with the 'unidirectionality' hypothesis (UDH, Sorace, 2004) which predicts deficits only with overt pronouns. Finally, although the potential role of processing cost is emphasized, the possibility that transfer may work in conjunction with processing (and both against native-like L2 performance) was not completely ruled out by the authors.

In a similar fashion, Belletti, Bennati, & Sorace (2007) examined the acquisition of 3rd person anaphoric subjects by English-speaking learners of Italian at near-native level of attainment[63]. As in the above-reviewed study of Sorace & Filiaci (2006), the authors used a PVT to test the interpretation of anaphors in forward and backward anaphora (such as the ones in examples (21) and (22) above). Additionally, they included a story-telling task in order to explore the production of null and overt pronouns. The results of both anaphoric interpretation and production tasks are totally in line with the findings of Sorace & Filiaci (2006). In the AR task, near-native speakers interpret overt pronouns as coreferential with the subject of the previous clause at a significantly higher rate than the native control group, whereas at the same time they overproduce overt pronouns in the story-telling task[64]. The authors, in contrast with Sorace & Filiaci (2006), do not interpret their results in terms of processing difficulties. More in line with the UA, they emphasize the effect of the L1 as the primary factor that hinders the performance of the near-native group. Crucially, this is one of the first formal SLA studies that highlights the need to explicitly consider the role of discourse in anaphoric production and interpretation. The

[63] Belletti et al. (2007) tested the acquisition of postverbal subjects as well. However, due to the scope of this thesis, this review focuses only on the tasks and results related to the acquisition of referential null and overt pronouns.

[64] It should be noted that the overproduction of overt pronouns reported in Belletti et al. (2007) was based on the simple counting of their proportion in the production data. Whereas the overall distributions of anaphoric forms may give us some hints regarding referential choices, they do not answer the crucial 'under what conditions' question (see also section 6.1). One group may use, overall, more or less overt pronouns than the other but this does not say much unless it is also determined whether these are pragmatically appropriate or not.

authors suggest that the licensing of null subjects in L2 is not a sufficient condition for the achievement of native-like competence: "any formulation of the null subject parameter has to be augmented by consideration of the discourse factors determining the distribution of syntactic options" (Belletti et al., 2007:682).

Whereas the 'processing vs representation' debate remains unresolved in the early syntax-discourse interface literature, Kras (2008) aimed to replicate the methodological design of the above-reviewed studies in order to test a crucial hypothesis: if the instability in the acquisition of anaphoric subjects is due to the influence of the L1, then Croatian learners of Italian[65], in contrast to the English-speaking learners tested in the previous studies, should exhibit native-like behaviour. The author claims that her SLA study in AR is the first to test speakers of a pro-drop L1 who are learning a pro-drop L2 and have reached near-native proficiency level (p.115). As in the previously reviewed studies, she used a PVT and tested the interpretation of null and overt subjects in ambiguous PAS constructions (such as the ones in examples (21) and (22) above). She found that the Croatian-speaking L2 learners, in contrast with the English-speaking learners in Sorace & Filiaci (2006) and Belletti et al. (2007), expressed native-like preferences with both null and overt subjects. Given that Croatian and Italian are assumed to be identical regarding the licensing and distribution of anaphoric subjects, the author contrasted her results to those of the aforementioned studies and concluded that "cross-linguistic influence is indeed an important (if not the main) cause of the instability at the discourse-syntax interface" (Kras, 2008:109). This conclusion runs against the strong version of the IH and is more in line with the UA and the IPH. It is also in line with the claims made in some early formal studies (Lozano, 2002b, 2002c; White, 1985, 1986) and runs against the findings of Bini (1993). Notice, however, that the few 'pro-drop L1/pro-drop L2' studies reviewed so far are not fully comparable. Apart from the different focus and the methodologies employed, the L2 learners examined in each study are of different proficiency levels. It could be the case that low-proficiency learners show deficits irrespectively of their L1, whereas more advanced learners perform differently.

[65] Croatian is a pro-drop language with "essentially identical discourse-pragmatic conditions for anaphora resolution as Italian" (Kras, 2008:109)

Turning our attention from L2 Italian to L2 Spanish, one of the first formal studies that explicitly considered the acquisition of the discourse/pragmatic properties of 3rd person anaphoric subjects (apart from their syntactic features) was Montrul & Rodríguez Louro (2006). The authors tested the production of null and overt subjects in L2 Spanish by English-speaking learners at three proficiency levels[66] (intermediate, advanced and near-native). The data of this study were elicited with oral retellings of the universally known story of 'Little Red Riding Hood'[67]. The authors considered first the overall distributions of null and overt subjects which demonstrated that intermediate learners produce significantly more overt subjects than any other group (the rest of the groups produce similar proportions of null and overt subjects). Additionally, the violation of discourse rules in the production of anaphoric subjects was explicitly considered in this study, since all items were tagged for being correct, redundant or illicit (as synonymous to ambiguous)[68]. The results revealed that intermediate and advanced learners produce more pragmatically redundant subjects than the native speakers, whereas the near-native learners do not differ from the Spanish control group. An example from an intermediate learners' production follows (Montrul & Rodríguez Louro, 2006:417):

---

[66] Overt subjects in the study of Montrul & Rodríguez Louro (2006) refer to both overt pronouns and noun phrases (merged under the label 'overt subjects'). This is probably the first time that lexical subjects are considered in a formal SLA study on the acquisition of the NSP. However, the inclusion of noun phrases in this study renders difficult the comparison with previous studies that only considered the binary 'null/overt pronoun' distinction.

[67] The authors underline the fact that isolated sentences as used in GJTs are not ideal for the study of discourse-constrained phenomena, such as anaphora, where "more than one or two connected sentences are needed" (Montrul & Rodríguez Louro, 2006:408). However, neither the elicitation of production data through the use of a universally known story is exempt of problems. As pointed out by Liceras, de la Fuente, & Sanz (2010), in the case of 'Little Red Riding Hood', the narrator may assume that the interlocutor is familiar with the story and may produce null subjects on the basis of this assumption.

[68] It should be noted, however, that the criteria used by the authors to determine the pragmatic felicity of the anaphoric subjects are bound to some subjective judgement. According to Montrul & Rodríguez Louro (2006:412), an overt subject is considered redundant when it is not used for 'emphasis'. A null subject is considered illicit when it is used for a 'switch of reference'. For the need of a more objective definition of redundancy and ambiguity see also 2.4.3.

23) **La Caperucita Roja$_i$** vive con su abuela$_j$. **La Caperucita$_i$** va al bosque. Un día **un lobo$_k$** va a la casa de la abuela$_j$ y **el lobo$_k$** come la abuela$_j$.

"**Little Red Riding Hood$_i$** lives with her grandmother$_j$. **Little Red Riding Hood$_i$** goes to the forest. One day **a wolf$_k$** goes to grandmother$_j$'s house and **the wolf$_k$** eats the grandmother$_j$"

Regarding the intermediate and advanced group, the results concerning redundancy are broadly in line with the findings of some early generative studies on the acquisition of the NSP in Spanish L2 by English-speaking learners[69] (Al-Kasey & Pérez-Leroux, 1998; Pérez-Leroux & Glass, 1997, 1999). The postulations of Stockwell et al. (1965) regarding the overuse of 3rd person subject pronouns by English learners of Spanish are also confirmed. On the other side, the largely unproblematic performance of the near-native group runs against the IH and other similar studies on Italian L2 (Belletti, Bennati, & Sorace, 2007; Sorace & Filiaci, 2006). Furthermore, Montrul & Rodríguez Louro (2006) found that the advanced and near-native learner groups occasionally produce some illicit null subjects (the differences with the native group bordering on significant). This is in line with the results of Lozano (2002b, 2002c) and Rothman (2007, 2009) who found that English-speaking learners overaccept illicit null subjects in CFC contexts. Finally, these results run against the UA (Sorace, 2004, 2006) which predicts that only overuse of overt pronouns should be expected. Overall, the non-target performance of the intermediate and advanced group is interpreted by the authors, against the processing accounts of the IH, in terms of potential L1 influence.

Lozano (2009b) is another L2 Spanish corpus-based study that tested the production of null and overt subjects (including noun phrases) by English-speaking learners of Spanish at two proficiency levels (advanced and upper-advanced). The author examined written narrative data from the CEDEL2 corpus and sought to determine whether the non-target performance of learners, reported in previous studies, concerns only specific grammatical persons or it affects the entire pronominal paradigm. Following a similar procedure with Montrul & Rodríguez Louro (2006), Lozano considered the pragmatic felicity of null and overt subjects in 'topic-shift' and 'topic-continuity' contexts. Learners (both advanced and upper-advanced) were found to produce more 3rd person redundant subjects than the

[69] Notice, however, that the results of studies within the scope of the syntax-discourse interface are difficult to compare with those of the early parametric studies, given that the latter focused exclusively on the grammaticality of the anaphoric subjects.

control group. An example from an advanced learner's production follows (Lozano, 2009b:152):

24) [Context: The informant is talking about the main character of the film "Spanglish"]

La madre$_i$ no puede hablar inglés pero **Ø$_i$** es muy trabajadora. **#Ella$_i$** empieza a trabajar… **#Ella$_i$** no puede comunicar[se] con esta familia…

"**The mother$_i$** cannot speak English but **(she)$_i$** is very hard-working. **#She$_i$** starts working… **#She$_i$** cannot communicate with the family…"

Crucially, the observed deficits concerned only animate referents and no differences were found between learners and native speakers regarding the 1st and 2nd grammatical persons. Additionally, some unpragmatic null subjects were occasionally found in the learners' data (against the UDH) although the differences with the natives did not reach significance with respect to this. The results confirmed the author's initial hypothesis that deficits are selective and do not affect the entire pronominal paradigm. Furthermore, regarding the (lower) advanced group, Lozano's findings are in line with Montrul & Rodríguez Louro (2006) and with earlier generative studies that briefly considered the pragmatic infelicity of anaphoric subjects (Al-Kasey & Pérez-Leroux, 1998; Pérez-Leroux & Glass, 1997, 1999). Regarding the upper-advanced group, the results are broadly in line with the corresponding studies on the interpretation of anaphoric subjects in Italian L2 by English-speaking learners (Belletti, Bennati, & Sorace, 2007; Sorace & Filiaci, 2006). The selectiveness of the deficits, however, led the author to postulate against both the representational and the processing accounts related to the IH. Lozano (2009b:161) concluded, instead, that "the observed deficits stem from the way Universal Grammar constrains pronominal features".

In a very recent study, Lozano (2016) focused only on very advanced English-speaking learners of Spanish and examined the production of anaphoric subjects in written narrative texts extracted from the CEDEL2 corpus. In line with his previous study (Lozano, 2009b), the author tested the pragmatic felicity of null and overt subjects (including noun phrases) in 'topic-shift' and 'topic-continuity' contexts. The results confirmed previous findings (Belletti, Bennati, & Sorace, 2007; Lozano, 2009b; Sorace & Filiaci, 2006) insofar as the very advanced English-speaking learners were found to produce redundant overt pronouns and noun phrases to a higher rate than native speakers. An example of a learner's production follows (Lozano, 2016:252):

25) Juno$_i$ es el personaje principal. #**Ella$_i$** vive con su padre$_j$ y su madrastra$_k$.

"Juno$_i$ is the main character. #**She$_i$** lives with her father$_j$ and her stepmother$_k$."

Crucially, this is the first SLA study on anaphora that considers the number of potential antecedents and their gender differences in the overexplicit behaviour of L2 learners. Lozano found that the production of redundant overt pronouns is triggered by the presence of more than one antecedent. Additionally, more noun phrases are produced in presence of three (or more) antecedents or two same-gender antecedents. The author proposed the Pragmatic Principles Violation Hypothesis (PPVH) according to which very advanced learners (as well as natives) tend to be more redundant than ambiguous in the production of anaphoric subjects due to the influence of pragmatic principles related to notions such as informativeness/economy and manner/clarity (Blackwell, 1998; Geluykens, 2013; Grice, 1975; Levinson, 1995).

Rothman (2009) tested the claim that the L2 acquisition of syntax-discourse interface phenomena such as anaphora is unattainable (Sorace & Serratrice, 2009; Valenzuela, 2006). The author tested the knowledge of the OPC[70] and the interpretation/production of null and overt pronouns by intermediate and advanced English-speaking learners of Spanish [71]. The pragmatically-constrained distribution of anaphoric subjects was examined in two experiments, namely a context felicitousness judgment and a translation task. In the first task, learners and native speakers were asked to judge the felicitousness of null and overt subjects in several contexts, such as in the following examples (Rothman, 2009:959):

26) (context supports overt subject)

My friends and I need newshoes. I always buy white shoes, but lately red shoes have become popular. I think about buying red shoes, but when we get to the shoe store I don't like any and decide to stick with my classic color. My friends still think the red shoes are really cool.

Ellos van a comprar los rojos y **yo** voy a comprar los blancos.

"They are going to buy the red ones and **I** the white ones."

[70] Since the present thesis focuses on referential null and overt subject expressions, the OPC task and results will not be reviewed here.

[71] Rothman (2009:957) argues that many of his 'advanced' learners are actually near-native speakers. However, due to the lack of objective measurements of proficiency, the authors opts to conservatively refer to them as 'advanced'. As already argued, there is a need for independent proficiency tests in SLA studies in order to ensure the comparability of the results.

27) (context supports null subject)

My girlfriend is studying abroad this semester. I'm very happy for her, but I miss her terribly. I really wish I were able to talk to her more.

Mi novia esta fuera del pais y **ø** nunca hablo con ella porque siempre esta ocupada.

"My girlfriend is out of the country and **(I)** never talk to her because she is always busy"

The intermediate learners, compared to the native speakers, failed to differentiate the appropriateness of anaphoric subjects according to pragmatic conditions. More specifically, they overaccepted both null and overt pronouns in contexts where the natives judged them to be pragmatically infelicitous. In the second task, the participants were asked to translate sentences from English to Spanish, such as in the example below (Rothman, 2009:960):

28) (the context being about a trip to Antarctica)

My dad starts asking me a million questions about the trip. He wants to know how I'm getting there, how long I'll be gone, if I have the proper equipment, etc. He asks me when the last time was I went to Antarctica and if Juan has any experience in the cold.

Translate: I've never visited there, but **he** has lived there for two years.

Crucially, in the Spanish version of the sentences, the selection of null or overt pronouns is constrained by the pragmatic conditions that were set according to the context (e.g. in CFC overt pronouns are expected). The intermediate learners were found to overuse both null and overt pronouns with respect to the native speakers. The fact that the deficits in L2 performance concern both types of subject expressions runs against the UDH. On the other hand, the very advanced groups behaved native-like in all tasks. This does not confirm the findings of comparable studies on Italian and Spanish L2 (Belletti, Bennati, & Sorace, 2007; Lozano, 2009b, 2016; Sorace & Filiaci, 2006) and runs against the strong version of the IH that predicts insuperable fossilization with the acquisition of anaphoric subjects. On the other hand, the results of the intermediate group are in line with the bulk of the previous literature that examined intermediate English-speaking learners of Spanish (Al-Kasey & Pérez-Leroux, 1998; Lozano, 2009b; Montrul & Rodríguez Louro, 2006; Pérez-Leroux & Glass, 1997, 1999). The interpretation of the results provided by Rothman considers at least three factors that may account for them. First, the effect of the L1: in line with the representational accounts, negative transfer from English is assumed to significantly constrain the anaphoric preferences of the learners. Second, the additional complexity involved in the interfaces may contribute to the delays in the acquisition of

anaphoric subjects. Third, learners are exposed to non-native input which derives from both the instructional context of SLA (teachers, peers) and the potential interaction with native speakers. In both cases, emphatic speech (technically redundant overt subjects) may be used as an additional aid in the communication with non-native speakers at lower levels of proficiency. This negative input, as also noted in Polio (1995), may act as a confound that promotes the overuse of overt subjects.

The role of input was also addressed in Keating, VanPatten, & Jegerski (2011) who contrasted advanced English-speaking learners of Spanish to heritage and native Spanish speakers. Under the assumption that the heritage speakers (Spanish-English bilinguals) have more consistent access to native Spanish input (due to early exposure), the authors set forth to test whether this would confer them an advantage with respect to L2 learners who have less access to native input. Both groups were contrasted to monolingual native speakers in the interpretation of ambiguous null and overt pronouns in PAS constructions, such as in the example below (Keating, VanPatten, & Jegerski, 2011:208):

29) Daniel$_i$ ya no ve a Miguel$_j$ desde que **Ø$_{i/j}$** se casó.
"Daniel$_i$ no longer sees Miguel$_j$ ever since **(he)$_{i/j}$** got married."

The results demonstrated that none of the two groups exhibited native-like preferences regarding antecedent assignment strategies. More specifically, bilinguals were found to overaccept the coreference of overt pronouns with subject antecedents (in line with the predictions of the IH), whereas null and overt pronouns were found to be in free variation for L2 learners. Crucially, these preferences were significantly different from the ones of the native speakers in the same task. The authors concluded that the role of input alone cannot account for the instability at the syntax-discourse interface. Alternatively, in a follow-up study (Jegerski, Keating, & VanPatten, 2011), the role of cross-linguistic influence as the source of instability was addressed. In line with the traditional experimental procedure, the authors examined the interpretative preferences of intermediate and advanced English-speaking learners of Spanish in PAS constructions. The novelty of this study, however, is that the role of discourse structure in terms of

semantic coordination and subordination relations was specifically addressed[72]. These relations are exemplified below (Jegerski, VanPatten, & Keating, 2011:489):

30) (coordination)

Jeffrey$_i$ saw Ricky$_j$ while **he$_{i/j}$** was hunting for coins in the fountain.

31) (subordination)

Anita$_i$ talked to her sister$_j$ after **she$_{i/j}$** had the baby.

Under the assumption that the selection of pronoun antecedents is discursively constrained in English (e.g. the ambiguous pronoun "he" in (30) would be more likely to refer back to the subject antecedent "Jeffrey" than would the "she" in (31) be likely to refer to "Anita"), whereas the processing of Spanish null and overt pronouns obeys principally to syntactic rules (such as those proposed in the PAS account), the authors examined how this cross-linguistic variation may be evidenced in the anaphoric preferences of L2 learners[73]. The results indicated that intermediate learners were not native-like in the interpretation of null and overt pronouns, whereas they did show evidence of L1 influence. This influence was also observed for the more advanced group which, however, exhibited some parallel native-like antecedent assignment strategy as well (only in coordinated discourse). The authors concluded that learners are not able to completely deactivate the L1 strategies and that "cross-linguistic influence occurs and may even be a primary cause of non-native behaviour, up through the advanced level" (Jegerski, Keating, & VanPatten, 2011:502).

A more recent study that considered the role of input in the acquisition of anaphoric subjects is Pladevall-Ballester (2013). The author used a GJT to examine the acceptability of null and overt pronouns in main and subordinate clauses by Spanish native speakers

[72] The authors draw upon the Segmented Discourse Representation Theory (SDRT; Asher, 1993) which considers the semantic relations between sentence constituents and makes predictions regarding the potential referents of linguistic anaphoric expressions (more details in Asher & Vieu, 2005). Note that, in SDRT, the terms 'coordination' and 'subordination' do not correspond to the traditional syntactically-determined coordinate and subordinate sentential relations.

[73] It should be noted that this assumption runs broadly against the bulk of literature on discourse anaphora outside the generative framework, where the relevance of syntactic and discursive factors is assumed to be universal (see Chapter 2).

and English-speaking learners of Spanish at three proficiency levels[74] (elementary, intermediate and advanced). An example from the linguistic items employed in this study follows (Pladevall-Ballester, 2013:129):

32) ¿Qué quieren hacer tus sobrinos$_i$? "What do your cousins want?"
a. **Ø$_i$** Quieren ir al parque. "(**They**)$_i$ want to go to the park"
b. #**Ellos$_i$** quieren ir al parque. "#**They$_i$** want to go to the park"

Regarding the elementary group, significant differences with the native speakers were found in the judgements of pragmatically (in)correct pronominal subjects. These differences concerned both main and subordinate clauses. The same picture was obtained for the intermediate group, although some development was also observed, insofar as the correct judgments increased with proficiency. The more advanced group was found to differ with the native speakers only with respect to pronominal subjects in subordinate clauses (differences in main clauses disappeared). The results of this study are broadly in line with the IH and relevant literature that documented persistent difficulties for English-speaking learners of pro-drop languages such as Italian or Spanish (Belletti, Bennati, & Sorace, 2007; Lozano, 2009b, 2016; Sorace & Filiaci, 2006). However, the author provided an explanation of the results in terms of the role of input. Taking into account the personal testimonies of learners who were asked to justify their judgments in the experimental tasks, Pladevall-Ballester considered the role of classroom instruction in the acquisition of anaphoric subjects. More specifically, learners expressed their awareness that null subjects are permitted in Spanish but, crucially, they showed incomplete knowledge of the discourse conditions that constrain their pragmatic felicity. In conclusion, the author claimed that her results broadly confirm the IPH and, in line with Rothman (2007), highlighted the importance of input in the acquisition of the discourse properties that constrain the interpretation and production of anaphoric subjects.

One of the scarce 'syntax-discourse interface' studies in the acquisition of anaphoric subjects that, similarly to Kras (2008), considers learners from an L1 background other than English is Margaza & Bel (2006). The authors tested the production of anaphoric

[74] The other NSP-related properties that were examined in Pladevall-Ballester (2013), namely expletives and postverbal subjects, are out of the scope of this thesis and shall not be reviewed here.

subjects[75] by intermediate and advanced[76] Greek-speaking learners of Spanish[77]. The first task employed in this study was a cloze production test where the participants had to fill in the missing subjects of sentences related to a previously-read narrative text. An example from the cloze task is provided below (Margaza & Bel, 2006:92):

33) ¿A dónde vas? … voy a Barcelona.
"Where are you going? … am going to Barcelona."

Additionally, the authors used a written production task where the participants were asked to describe a difficult situation of their life. Results from the first task (overall distributions) revealed that intermediate learners overuse overt subjects whereas the advanced participants almost reached the percentage of the native speakers in the production of null subjects. Regarding the free production task, once again the intermediate students overused pronouns in cases where the expression of pronominal subjects is considered redundant, such as in the example below (Margaza & Bel, 2006:95):

34) #**Yo** no voy con el coche y con la motocicleta porque Ø no quiero estar muerto.
"#**I** don't go with the car and with the bike because (I) don't want to die."

On the other hand, the advanced Greek-speaking learners were not found to overuse anaphoric subjects. The attested overproduction of pronouns by the intermediate Greek-speaking learners is fully in line with the findings of Bini (1993). The authors interpreted these results, in the same line with the aforementioned study, in terms of an SLA strategy related to problems on the acquisition of inflectional morphology. The fact that the advanced informants did not overuse pronominal subjects confirmed the original hypothesis of Margaza & Bel (2006) that the competence level affects the production of anaphoric subjects. Most importantly, although the learner participants in this study have not reached near-native levels of proficiency, this finding is in line with Kras (2008) who

---

[75] Margaza & Bel (2006) also examined the acquisition of the second property attributed to the NSP, namely the subject inversion, which is not being considered in this thesis. Therefore, we focus exclusively on the results that concern referential null and overt subjects.

[76] The proficiency level of the learners was determined by the hours of exposure to Spanish. For the methodological problems related to the lack of objective proficiency measurements see 5.2.

[77] As already discussed, Greek is a pro-drop language like Spanish (see section 3.1).

suggested that the lingering deficits observed in previous studies on the acquisition of anaphoric subjects by English-speaking learners are mostly due to L1 influence. The results, overall, run against the strong version of the IH and partially support the representational accounts (UA and IPH). However, the non-target behaviour of the intermediate learners led the authors to conclude that transfer plays an important role but may not be the only relevant factor in the acquisition of anaphoric subject expressions.

A more recent study that examines non-Anglophone learners of L2 Spanish is García-Alcaraz & Bel (2011). More specifically, the authors tested the anaphoric production of native Spanish speakers and L2 learners with Moroccan Arabic (MA) L1 in PAS structures extracted from elicited oral and written narrative data [78]. Following a methodological design previously employed in Berman (2008), participants were shown a short silent film about a conflict situation and they were later asked to narrate a similar personal experience. Following Bel, Perera, & Salas (2010), the authors annotated the produced 3rd person subject expressions [79] (null, overt pronouns, noun phrases [80]) according to discourse function (introduction, reintroduction, maintenance) and syntactic function of the antecedent (subject, direct object, indirect object, other). Regarding the syntactic function of the antecedent, no significant differences were found between learners and native speakers. Both groups employ null pronouns for subject antecedents (as predicted by the PAS) and overt pronouns for subject antecedents as well (against the PAS). Regarding the discourse function of the subject expressions, learners were found

[78] The proficiency level of the L2 learners in this study is not specified. According to the authors (García-Alcaraz & Bel, 2011:170), the participants have been studying Spanish and Catalan in high school for the last three years. It is reasonable to assume that their proficiency level in Spanish could be somewhere between intermediate and advanced.

[79] Due to the nature of the elicited narrations ('personal experience'), we might expect mostly 1st person subject expressions to be produced. Indeed, the authors acknowledge that the number of 3rd person anaphors is limited in their data (García-Alcaraz & Bel, 2011:171). No raw frequencies of the analysed items were provided by the authors.

[80] Although lexical subjects are tagged in this study, they are not taken into consideration during the analysis of the data. The reasons for this decision are not made explicit. Given that the annotation scheme follows Bel et al. (2010), it might be due to the assumption held in the aforementioned study that "NPs have, by definition, no antecedent"(p.246). It is not clear to me, however, why noun phrases may be assumed not to have antecedent.

to overuse overt pronouns for reference maintenance, such as in the example below (García-Alcaraz & Bel, 2011:174):

35) ella$_i$ se rebota ("she$_i$ gets angry")
y Ø$_i$ se piensa ("and (she)$_i$ thinks")
que es de verdad ("that it is true")
y #**ella$_i$** también pues comienza a insultar
("and #**she$_i$** also starts to insult")

This redundant linguistic behaviour of low-proficiency learners in a language pairing where both source and target language are pro-drop is in line with the findings of Margaza & Bel (2006) and Bini (1993). Furthermore, it broadly supports the processing accounts of the IH framework (Sorace & Filiaci, 2006; Sorace, Serratrice, Filiaci, & Baldo, 2009). Given the similarity between L1 and L2, the role of negative cross-linguistic influence may be discarded in this case. Accordingly, the authors provided an explanation of the results in terms of the processing difficulties related to the syntax-discourse interface.

In a more recent study (Bel & García-Alcaraz, 2015), the authors used AJTs to examine the acceptability of null and overt pronouns in PAS structures by native speakers and intermediate MA learners of Spanish. As a novelty of this study, the implicit causality of the verbs was supposedly controlled in order to ensure that the experimental sentences were completely ambiguous[81]. Additionally, clause order was manipulated in the AJTs (main-subordinate and subordinate-main clause orders were separately tested). Examples of two of the sample items for main-coordinate clause order follow (Bel & García-Alcaraz, 2015:217):

36) (Main-subordinate clause. Null pronoun. Subject antecedent.)
Iker$_i$ evita a Iván$_j$ cuando **Ø$_{i/j}$** tiene problemas. Iker tiene problemas.
"Iker$_i$ avoids Ivan$_j$ when **(he)$_{i/j}$** has problems. Iker has problems."

37) (Main-subordinate clause. Null pronoun. Object antecedent.)
Ángel$_i$ asustó a Héctor$_j$ mientras **Ø$_{i/j}$** entraba en la habitación. Héctor entraba en la habitación.

[81] It should be noted, however, that the claims made in the literature regarding the implicit causality of verbs concern specifically causal constructions (Caramazza, Grober, Garvey, & Yates, 1977; Garvey & Caramazza, 1974; Goikoetxea, Pascual, & Acha, 2008; Hartshorne, Sudo, & Uruwashi, 2013; McKoon, Greene, & Ratcliff, 1993) whereas none of the examples included in the sample of stimulus items provided by Bel & García-Alcaraz (2015:217) is of this type.

"Angel$_i$ scared Héctor$_j$ while **Ø$_{i/j}$** came in the room. Hector came in the room."

Two observation can be made regarding the sample items in (36) and (37) and the PAS structures in general. Firstly, even in experimental settings, it is difficult to create globally ambiguous contexts. In fact, it may be argued that in (36) it is "Ivan" who is more likely to have problems (based on a world-knowledge-based assumption that a person A is more likely to avoid a person B who has problems than the other way round). Similarly, in (37), the person that enters the room is more likely to be scared by the person that is already in the room. Secondly, even if the creation of completely ambiguous contexts is assumed to be achievable, it would have little or no correspondence with real discourse, where ambiguity is practically inexistent (see section 2.4.3). Whatever the case, and regarding the results of Bel & García-Alcaraz (2015), learners were found to significantly differ from the native speakers only in the subordinate-main clause constructions. More specifically, differences were found in the 'null pronoun to subject' and 'overt pronoun to object' conditions. In the former condition, native speakers accepted almost categorically the coreference of a null subject with the subject of the previous clause, whereas in the latter condition they were equally categorical in their acceptance of an overt pronoun as being coreferential with the object of the previous clause. Learners were less categorical in their acceptability rates in both conditions. These results are broadly in line with previous studies focusing on low-proficiency learners of Spanish with a pro-drop L1 background (García-Alcaraz & Bel, 2011; Margaza & Bel, 2006). The authors ruled out the role of transfer and provided a processing explanation of the results, in line with the IH. Additionally, in line with other previous studies (Pladevall-Ballester, 2013; Rothman, 2009), they highlighted the importance of input with respect to the phenomenon under study. The lack of formal instruction regarding the discourse properties of anaphors and the underrepresentation of overt pronouns in native Spanish may act as conflating factors which hinder the successful acquisition of anaphoric subjects.

Another recent study that examined the acquisition of anaphoric subjects in Spanish L2 by learners with pro-drop L1 background is Judy (2015). The author used both offline and online tasks to test native speakers and Farsi-speaking learners of Spanish at near-

native level of proficiency[82]. First, their corresponding ratings regarding the pragmatic suitability of null and overt pronouns in three contexts (topic maintenance, topic shift and contrastive focus) were compared through an offline context-matching felicitousness task (CMFT). The three experimental contexts are exemplified below (Judy, 2015:178-179):

38) (Topic maintenance)

Mi cuñada$_i$ es muy sociable. Ø$_i$ Tiene muchos amigos y por eso Ø$_i$ va a muchas cenas a la canasta donde Ø$_i$ tiene que contribuir con algo.

"My daughter-in-law$_i$ is very social. (She)$_i$ has a lot of friends and for that reasons, (she)$_i$ goes to a lot of potluck dinners where (she)$_i$ has to share something."

a. Así que #**ella$_i$** lleva postres y Ø$_i$ comparte todo con sus amigos.

b. Así que **Ø$_i$** lleva postres y Ø$_i$ comparte todo con sus amigos.

"So, **she$_i$** takes desserts and (she)$_i$ shares everything with her friends."

39) (Topic shift)

Mi hija$_i$ quiere ser autora y Ø$_i$ no tiene otros intereses. Yo$_j$ creo que es mejor tener varios intereses y Ø$_j$ sugiero otras actividades, pero no importa lo que diga yo$_j$.

"My daughter$_i$ wants to be an author and (she)$_i$ has no other interests. I$_j$ think that it is best to have various interests and (I)$_j$ suggest other activities, but it doesn't matter what I$_j$ say."

a. Finalmente **ella$_i$** escribe cuentos y Ø$_i$ pasa todo el día en su cuarto.

b. Finalmente #**Ø$_i$** escribe cuentos y Ø$_i$ pasa todo el día en su cuarto.

"In the end, **she$_i$** writes stories and (she)$_i$ spends the whole day in her room."

40) (Contrastive focus)

Cuando salimos a cenar, mi novia$_i$ prefiere comer platos livianos, pero yo$_j$ prefiero comer algo sustancioso.

"When we go out to eat, my girlfriend$_i$ prefers to eat light dishes, but I$_j$ prefer to eat something of substance."

a. Así que **ella$_i$** come ensaladas y **yo$_j$** como milanesas en los restaurantes.

b. Así que #**Ø$_i$** come ensaladas y #**Ø$_j$** como milanesas en los restaurantes.

"So, **she$_i$** eats salads and **I$_j$** eat breaded meats in restaurants."

[82] Judy (2015) is probably the first SLA study in the acquisition of Spanish anaphoric subjects that uses online methodology to test the claim that processing difficulties are inherent to the syntax-discourse interface. Whereas this is undoubtedly useful and may provide some novel evidence regarding the IH, it should not be considered as the only valid methodological design for this purpose (Sorace, 2011:20).

In the offline CMFT, learners were found to differ from the native speakers in some contexts. More specifically, they tolerated, to a significantly higher degree than Spanish natives, ambiguous null subjects in topic-shift and redundant overt pronouns in topic-maintenance contexts. This is in line with the findings of Lozano (2009b, 2016) regarding upper-advanced Spanish learners and may be interpreted as running in favour of the IH and some relevant studies in L2 Italian (Belletti, Bennati, & Sorace, 2007; Sorace & Filiaci, 2006). On the other hand, these offline-data results run against Kras (2008) who found native-like L2 behaviour in a similar language pairing (L1 pro-drop/L2 pro-drop). Judy employed, additionally, an online self-paced reading task (SPRT) in order to determine if the learners' processing differed from that of the native speakers. According to the author, the L2 groups showed native-like processing in all contexts[83]. Crucially, in the topic-shift contexts (considered particularly problematic in the IH), learners processed target sentences in a native-like manner. This finding runs sharply against the processing-deficit account of the IH.

Another recent study that employed both offline and online methodology is Bel, Sagarra, Comínguez, & García-Alcaraz (2016). The authors contrasted L2 Spanish learners of two different L1 backgrounds (English and Moroccan Arabic) at three proficiency levels (intermediate, upper-intermediate, advanced) in comparison with a native Spanish control group. As the authors note (p.154), this methodological setting, which was also used in some early parametric studies (Lozano, 2002b, 2002c; White, 1985, 1986), is ideal for examining potential cross-linguistic influence. Crucially, Bel and colleagues compared the AR preferences (in L2 Spanish) of learners with pro-drop L1 (Moroccan Arabic) and learners with non-pro-drop L1 (English) under the same conditions. First, an SPRT was used to measure reading times (RTs) in the following four PAS-like conditions of backward anaphora[84] (Bel, Sagarra, Comínguez, & García-Alcaraz, 2016:146):

---

[83] Note, however, that the native speakers were faster than the learners in all contexts. Additionally, they did not detect the (assumed by the author) infelicitous referential expressions in several contexts, since their reaction times were not slowed down in these contexts, as expected. This may entail some methodological problems regarding the appropriateness of the measures, since the native speakers were found to process very fast independently of whether the sentence was pragmatically correct or not.

[84] Note that the authors use the term 'backward anaphora' as defined in Tsimpli et al. (2004) and reversely to the definition used in Sorace & Filiaci (2006). Regarding the inconsistent use of the terms 'backward' and 'forward' anaphora in the literature see also footnote 61.

41) Condition 1: Null subject pronoun -- Subject antecedent

El músico$_i$ saluda al bombero$_j$ mientras **Ø$_i$** lleva un violín en la mochila.

"The musician$_i$ greets the fireman$_j$ as **(he)$_i$** carries a violin in his backpack."

Condition 2: Overt subject pronoun -- Object antecedent

El músico$_i$ saluda al bombero$_j$ mientras **él$_j$** lleva un casco en la mochila.

"The musician$_i$ greets the fireman$_j$ as **he$_j$** carries a helmet in his backpack."

Condition 3: Null subject pronoun -- Object antecedent

El músico$_i$ saluda al bombero$_j$ mientras **Ø$_j$** lleva un casco en la mochila.

"The musician$_i$ greets the fireman$_j$ as **(he)$_j$** carries a helmet in his backpack."

Condition 4: Overt subject pronoun -- Subject antecedent

El músico$_i$ saluda al bombero$_j$ mientras **él$_i$** lleva un violín en la mochila.

"The musician$_i$ greets the fireman$_j$ as **he$_i$** carries a violin in his backpack."

The online experiment was followed by an offline procedure in which the participants had to answer questions aimed to assess preference-biased interpretations. Regarding the online data, both advanced learner groups were native-like. Differences were found only between the native speakers and the lower-proficiency learners of both groups. On the other hand, the offline data showed that the advanced Arabic-speaking participants, in contrast to their English-speaking counterparts, fully converged with the Spanish natives at the interpretative level of AR. This finding is broadly in line with White (1985, 1986) and Lozano (2002b, 2002c). It further supports the claims of Kras (2008) with respect to the role of transfer. Overall, the results indicate a developmental progression which, with the exception of the English-speaking groups in the offline task, finally leads to native-like attainment. In line with other studies (Judy, 2015; Kras, 2008; Rothman, 2009) the results confirm that deficits in the syntax-discourse interface are not insurmountable, against the claims of the processing accounts related to the IH (Sorace, 2011; Sorace & Serratrice, 2009; Valenzuela, 2006).

Very recently, Lozano (forthcoming) examined Greek-speaking learners of Spanish at three proficiency levels (intermediate, lower-advanced and upper-advanced). Under the assumption that AR behaves similarly in Spanish and Greek, the author set forth to investigate whether the L1 may be a facilitating factor in the acquisition of anaphoric subjects. Learners were tested through an AJT in three contexts (topic-continuity,

contrastive focus and emphatic) and their acceptability rates were compared to those of a native Spanish control group. Examples of the three aforementioned contexts are provided below:

42) (Topic-continuity)

Diego$_i$ tiene mucho dinero aunque **Ø$_i$/#él$_i$** trabaja poco.

"Diego$_i$ has a lot of money although **(he)$_i$/#he$_i$** works a little."

(Contrastive focus)

Aunque Michael Douglas$_i$ y Sharon Stone$_j$ ganan muchos millones al año, **el$_i$/ella$_j$/#Ø$_{i/j}$** trabaja poco.

"Although Michael Douglas$_i$ and Sharon Stone$_j$ earn many millions per year, **he$_i$/she$_j$/#Ø$_{i/j}$** works a little."

(Emphatic)

En el banco ha desaparecido una suma importante de dinero. El director del banco sospecha de sus empleados, Roberto$_i$, Alfonso$_j$ y Manuel$_k$, aunque Alfonso$_j$ afirma que **el$_j$/#Ø$_{i/j/k}$** no tiene el dinero.

"An important amount of money has disappeared at the bank. The bank director suspects his three employees, Roberto$_i$, Alfonso$_j$ and Manuel$_k$ of robbery, although Alfonso$_j$ claims that **he$_j$/#(he)$_{i/j/k}$** doesn't have the money."

The author, in line with Jegerski et al. (2011), highlights the fact that the PAS structures, which have been broadly used in previous experimental research, may not reflect the complexity of anaphora in real discourse. Lozano seeks to overcome this limitation by including contrastive focus and emphatic contexts in the experimental design of his study, as can be seen in (42). The results indicate that not all anaphoric patterns are equally problematic in L2 acquisition. Upper-advanced learners behaved native-like in contrastive contexts, whereas they overaccepted redundant overt pronouns in topic-continuity contexts. Intermediate and lower-advanced groups were found to significantly differ from native speakers regarding both contexts, although the differences become less pronounced as proficiency grows. Finally, some optionality was observed for all groups (both natives and learners) in the emphatic contexts. Regarding the contrastive focus experiment, the results run against the claim of the IH that all syntax-discourse properties are inherently problematic (Sorace, 2011; Sorace & Serratrice, 2009; Valenzuela, 2006). On the other hand, given that cross-linguistic influence is discarded, the overexplicit behaviour of the upper-advanced learners in topic-continuity contexts provides some evidence that the L1 may not be a facilitating factor in the acquisition of anaphoric subjects (in line with: Bini, 1993; Margaza & Bel, 2006; Polio, 1995 and against: Bel et al., 2016; Kras, 2008; Lozano, 2002b, 2002c; White, 1985, 1986). The author provided an interpretation of the results in terms of the pragmatic principles of economy recovered

in the PPVH (Lozano, 2016). According to this hypothesis, learners prefer to be redundant instead of ambiguous, because ambiguity (in contrast to redundancy) may lead to a communication breakdown.

3.2.6 Summary of results of the 'syntax-discourse interface' literature

The research discussed in the previous section provides a considerable amount of mixed evidence. Given the differences in the methodology, the examined populations and the specific focus of interest of each study, an overall comparison of the results may be considered risky and untenable[85]. It is crucial, however, to highlight some important points of consensus in order to move forward. First, L2 learners from different L1 backgrounds (both pro-drop and non-pro-drop) show deficits in some features related to the interpretation and/or distribution of anaphoric subjects in pro-drop languages such as Italian, Spanish and Greek. This consensual finding implies that something else, other than transfer, may be also involved in the acquisition of anaphoric subjects. This has led many authors to postulate other explanations for the non-target behaviour of L2 learners (lack of processing resources, input-related factors, universal pragmatic principles, etc.). An issue that certainly remains unresolved is whether these deficits are persistent (insuperable) or not. On one hand, the results for low-proficiency L2ers are consistent and show deficits in the bulk of the literature. On the other hand, highly advanced and near-native learners have been found to perform in a native-like fashion in some studies (but not in others). Second, and in relation to the first, there is compelling evidence that deficits mostly concern the overproduction/overacceptance of overt subjects (redundancy). In the opposite direction, some problems with null subjects (ambiguity) have also been reported, though to a lesser degree. Thus, the issue regarding the directionality of deficits remains partly unresolved. Third, the deficits have been observed in offline interpretation/production tasks but, crucially, not with online data. This may be,

[85] To just mention a few methodological differences: (i) most studies focus exclusively on null and overt pronouns whereas some studies consider the entire set of referential subject expressions (including lexical subjects), (ii) given the lack of standardized measures, the advanced proficiency level of the learners in some studies may correspond to the upper-advanced, lower-advanced or even near-native level in other studies, (iii) most studies focus on the interpretation of null and overt pronouns in PAS structures, whereas some studies include contrastive focus contexts (a few corpus-based studies even consider the entire set of AR patterns in real discourse).

partly, due to the fact that the bulk of the studies have traditionally employed offline methodologies and online data are still scarce. Whatever the case, the fact that no deficits have been found in the few online studies conducted so far runs against the processing accounts of the IH. Fourth, there is a broad consensus in the interfaces-related literature that the role of discourse/pragmatics is crucial for anaphora. By definition, the syntax-discourse interface encompasses cognitive aspects outside pure syntax. However, and despite this consensual assumption, most anaphora studies in SLA have relied on experimental data to test the syntactically-driven interpretation of null and overt pronouns in PAS structures. The entire referential paradigm and the ensemble of anaphoric patterns have been rarely considered in formal/generative approaches (with the exception of some corpus-based studies).

Due to the particular methodological setting of the present study, special attention should be given to the findings of studies that examine similar populations as the ones examined here (i.e. both the source and the target language are pro-drop). Overall, the findings of these studies may be summarized as follows:

i. Low proficiency L2 learners with pro-drop L1 background have been consistently found to perform in a target-deviant fashion in the L2, despite the similarity between source and target language. More specifically, the intermediate Greek-speaking learners of L2 Spanish in Margaza & Bel (2006), the intermediate Spanish-speaking learners of L2 Italian in Bini (1993) and the intermediate Arabic-speaking learners of L2 Spanish in Bel & García-Alcaraz (2015) were found to accept and/or produce redundant overt pronouns. As already argued, this indicates that transfer cannot be the only cause of non-target performance.

ii. On the other hand, the evidence regarding more proficient learners is less consistent. The near-native Croatian-speaking learners of L2 Italian in Kras (2008), the advanced Greek-speaking learners of L2 Spanish in Margaza & Bel (2006) and the advanced Arabic-speaking learners of L2 Spanish in Bel, Sagarra, Comínguez, & García-Alcaraz (2016) were found to perform in a native-like fashion. However, the near-native Farsi-speaking learners of L2 Spanish in Judy (2015) and the upper-advanced Greek-speaking learners of L2 Spanish in Lozano (forthcoming) showed persistent deficits. Thus, even for L2ers whose L1 may be a facilitating factor, the possibility of ultimate attainment in the acquisition of anaphoric subjects remains an unresolved issue.

iii. Finally, it should be noted that, in the studies where learners from pro-drop L1 background were contrasted with learners from non-pro-drop L1 background, the former were always found to perform better than the latter (Bel, Sagarra, Comínguez, & García-Alcaraz, 2016; Fernandez de Moya, 1996; Lozano, 2002b, 2002c, 2008b, White, 1985, 1986). This provides evidence in favour of the role of the L1 as a facilitating factor.

Outside the formal/generative framework, some research on the L2 acquisition of anaphoric subjects has been recently conducted under variationist and discourse/pragmatic approaches. The non-generative SLA studies that have focused on this issue shall be reviewed in the next section.

3.3 Discourse-oriented approaches on the acquisition of anaphoric subjects

Outside the generative framework, anaphora has been extensively studied from variationist/probabilistic and discourse/pragmatic perspectives[86]. As already discussed, several theoretical models of discourse anaphora (Ariel, 1990; Givón, 1983; Gundel, Hedberg, & Zacharski, 1993; Kibrik, 2011) have been proposed in order to account for the interpretation and distribution of anaphors in native discourse (see section 2.4). However, these models have been seldom applied in the domain of SLA which is largely dominated by formal/generative approaches. One important reason for this may be that the straightforward application of discourse-oriented approaches for the exploration of L2 data does not seem fully adequate. Polio (1995), one of the first discourse-oriented SLA studies on anaphoric subjects, examined the production of zero pronouns in the interlanguage of English-speaking learners of Chinese. The author made a fundamental

[86] Despite their differences in theoretical aspects and in the corresponding analyses they adopt, discourse/pragmatics and variationist approaches will be examined in the same main section under the label of 'discourse-oriented approaches'. The reason for this decision is that they share two fundamental characteristics, in opposition to formal/generative approaches. First, they aim to examine the full range of anaphoric forms and second, they specifically consider the role of contextual factors in cognitive and/or discursive terms (Quesada, 2015:206). Additionally, some corpus-based studies that were reviewed in section 3.2 also share a lot with discourse-oriented approaches (e.g. Lozano, 2009b, 2016). As already argued in the introduction of Chapter 3, theoretical approaches are not categorical in nature.

observation as she considered the possibility of following a discourse-oriented analysis previously employed in Williams (1988)[87]:

> The results would probably show that in Chinese interlanguage zero pronouns, lexical pronouns, and full NPs are used in the same way, relative to each other, as they are in native Chinese, English, and other languages. In the future it might be worthwhile to complete this analysis, but in the end it probably would not tell us why (or how) the NNSs were overusing pronouns.

In other words, the straightforward application of discourse models of anaphora to L2 data may provide useful insights regarding several aspects. It may tell us whether the referential choices of native speakers and learners are constrained by the same discourse factors. If so, it may further tell us whether these (assumed to be universal) discourse constraints operate to the same degree in L1 and L2. In that sense, it may broadly provide evidence of an overall target-like or target-deviant performance. However, a more fine-grained consideration of the specific contexts in which non-native speakers experience difficulties (in terms of overuse and underuse) is needed in order to answer how and why they do so (if this is the case). Bearing this in mind, an overview of the discourse-oriented studies on the acquisition of anaphoric subjects in L2 Spanish will be performed in the following section.

3.3.1 Variationist studies on the acquisition of Spanish anaphoric subjects

A considerable amount of studies on SLA have been conducted under the variationist framework (also known as variationist sociolinguistics[88]). Variationist approaches draw on the seminal work of Labov (1966, 1972) who proposed a theoretical model that

---

[87] Williams (1988) focused on English L1/L2 production data and coded each referential expression in relation to Givón's (1983) categories. Subsequently, the author compared the proportions of null subjects in the data of native speakers and learners for each category and found that the general discourse function for zero anaphora is similar across L1 and L2 groups.

[88] The label 'sociolinguistics' has been the source of some misconceptions regarding the aims of variationist approaches. Although it might lead to assume that variationist sociolinguistics is particularly interested in social aspects, this is not entirely accurate: "In many variationist studies, the list of factors which determine probabilities of occurrence includes no so-called social features" (Bayley & Preston, 1996:26).

revolves around the notion of linguistic variation. According to this model, linguistic forms are inherently variable due to the influence of a broad range of (socio)linguistic factors. Practitioners of this field seek to uncover, by means of sophisticated quantitative analysis of linguistic data (usually elicited through oral interviews), the nature of these interacting factors. Essentially, under the variationist framework, linguistic patterns are assumed to be probabilistically determined. Regarding SLA, several authors sought early on to extend the variationist model to account for L2 acquisition[89] (Bayley & Preston, 1996; Ellis, 1985; Gass, Madden, Preston, & Selinker, 1989; Tarone, 1988). The alternation of anaphoric forms in discourse seems, a priori, to be ideal for the investigation purposes of variationist approaches, given that variation aims to account for phenomena such as anaphora that may entail "different ways of saying the same thing" (Labov, 1969:72). Nevertheless, very little variationist research has been carried out in the specific field of the L2 acquisition of Spanish anaphoric subjects[90]. Crucially, from the perspective of variationist sociolinguistics, the alternation of anaphoric subject expressions in discourse is assumed to be "the result of competing factors which contribute their relative weight to the likelihood that a pronoun will or will not be used in a variable context" (Shin & Erker, 2015:171).

Geeslin & Gudmestad (2008) is probably the first variationist study on the L2 acquisition of Spanish anaphoric subjects. The authors examined the relevance of two factors (the

[89] Variationist approaches broadly reject the independence of a purely linguistic competence in terms of UG constraints, as proposed by the generative paradigm, and argue instead in favour of a communication-oriented model of language (Ellis, 1989:42). The differences between the two frameworks have been the reason for a lot of tension between scholars from each field (Bayley & Preston, 1996:20; Pérez-Leroux & Glass, 1999:220; Quesada, 2015:207).

[90] By contrast, there is abundant literature on anaphoric subjects in native Spanish (childhood acquisition , attrition and Spanish L1) from variationist perspectives (Bayley & Pease-Alvarez, 1997; Bentivoglio, 1983, 1987; Cameron, 1995; Cameron & Flores-Ferrán, 2004; Dumont, 2006; Flores-Ferrán, 2002, 2004; Hurtado, 2005; Otheguy & Zentella, 2012; Shin & Erker, 2015; Shin & Montes-Alcalá, 2014; Shin & Otheguy, 2009; Silva Corvalán, 1982, 1994; Travis, 2007; Travis & Cacoullos, 2012).

specificity of the referent and the grammatical person/number[91]) for the use of different anaphoric subject expressions in oral interview-elicited production data from native speakers and very advanced English-speaking learners of Spanish. It should be noted that this is the first study to consider the entire paradigm of grammatical persons and the full range of subject forms in Spanish L2 (null/overt pronouns, lexical noun phrases, demonstrative/interrogative/indefinite pronouns). The overall distributions of anaphoric subjects revealed subtle differences between learners and native speakers for some grammatical persons and no differences for others[92]. More specifically, learners used overall more null subjects than the native speakers (especially in 2nd person singular/plural) but, in contrast, they produced more noun phrases in 3rd person singular forms[93]. This finding is in line with the results in Lozano (2009b, 2016). The authors concluded that subject expression in L2 is "tremendously complex" and that the interaction of person, number and specificity of the referent is important for the complete description of the phenomenon. They also highlighted the fact that differences between L2ers and native speakers may exist outside the, traditionally examined, binary null/overt pronoun distinction and suggested the adoption of a broader scope. In a follow-up study, Gudmestad & Geeslin (2010) used the same methodology and examined the same dataset of the previous study in order to test the effect of two additional factors

[91] The specificity of the referent "refers to whether or not the referent is a clearly identifiable entity or an unspecified group or individual" (Geeslin & Gudmestad, 2008:71). Results regarding this factor are out of the scope of this thesis and shall not be thoroughly examined here.

[92] It should be noted that, in this and follow-up studies (Geeslin & Gudmestad, 2011; Gudmestad & Geeslin, 2010), first and subsequent mentions of a referent were not distinguished. It is reasonable to assume, however, that the initial mention of a referent in discourse may be strongly biased against the selection of null forms (more on this in section 6.1). Under this assumption, this potentially categorical context is separately considered in more recent studies (Gudmestad, House, & Geeslin, 2013).

[93] As the authors acknowledge, "we do not wish to imply that the use of these null subjects on the part of our learners is inappropriate. In fact, no evaluative assessment has been made. It is quite possible that NNSs simply produced different types of discourse in which more null subjects are possible" (Geeslin & Gudmestad, 2008:81). However, as already discussed, it is always crucial to determine the appropriateness of referential choices (in pragmatic terms) in order to account for potential non-target performance.

(tense/mood/aspect (TMA) of the verb and switch-reference[94]) in the production of anaphoric subjects. The authors found that native speakers and very advanced L2 learners exhibited more similarities to each other than differences. More specifically, both groups used significantly more null than overt subjects for same-reference contexts. This finding was further confirmed in a third study (Geeslin & Gudmestad, 2011), where the additional factors of referent cohesiveness[95] and perseveration[96] were examined for the same participant pool and dataset. General patterns for native speakers and L2ers were found to be similar, insofar as more null subjects were used in same-reference contexts and/or after another null subject in the previous clause (in contrast, more overt subjects were used in switch-reference and after overt subjects). The authors concluded that the overall rates of use of anaphoric subjects vary in the same direction for native speakers and learners across the categories examined. Finally, they hypothesized that the higher overall rates of null subjects in the production of learners, also reported in their previous studies, may have two possible explanations. Either learners produce an inherently simpler discourse that allows greater use of null subjects or they allow more ambiguity in their speech than native speakers. Whatever the case, no answer may be given to this empirical question unless the specific discourse structures where null subjects appear are examined.

Abreu (2009), in a similar fashion with the above-reviewed studies, compared native speakers of Spanish to Spanish/English bilinguals and English-speaking learners of L2 Spanish[97]. The author examined the production of the three aforementioned groups in oral

---

[94] The switch-reference variable was aimed to identify "whether the subject of the preceding finite verb was the same as the subject of the current token" (Gudmestad & Geeslin, 2010:275). For more information on this factor see 6.2.2.

[95] Referent cohesiveness is a more fine-grained version of the switch-reference variable that includes nine (instead of two) subcategories. We remit to the study where it was initially proposed for more information (Silva, 1993). Note that some of these nine subcategories are exclusively applicable to interview data and may not account for other discourse genres (e.g. written narratives) (Geeslin & Gudmestad, 2011:21).

[96] Perseveration (also known as 'priming') examines whether an overt pronoun is followed by an overt pronoun (and, inversely, whether a null subject is followed by null subject) in discourse (Cameron, 1994; Cameron & Flores-Ferrán, 2004).

[97] The proficiency level of the learners in this study was determined through self-reported ratings. On a scale from 1 to 7, seven (out of ten) L2ers classified their Spanish proficiency as 5, while one chose 4 and

interview-elicited data. Crucially, all grammatical persons were merged and noun phrases were not considered in this study. The overall distribution showed that L2ers used significantly more null subjects (for all grammatical persons together) than native speakers and bilinguals. However, a closer look at the 3rd person singular subjects reveals that the English-speaking learners employ overt Spanish pronouns almost twice as much as the native speakers (56% vs 31%). This finding is in line with the results in the bulk of the L2 corpus-based formal/generative literature (Lozano, 2009b, 2016; Montrul & Rodríguez Louro, 2006; *inter alia*). The results, taken together, highlight the need for the separate examination of each grammatical person. Regarding the factors that constrain the referential choices of each group, the author concluded that "L2 learners are able to acquire the most important constraints that also influence monolingual pronominal production" (p.177).

In a more recent study, Linford & Shin (2013) examined whether lexical frequency of verb forms may influence the production of anaphoric subjects in L2 Spanish. The authors compared the oral interview-elicited data of two English-speaking groups of learners (intermediate and advanced proficiency levels). It should be noted that there was no native control group in this study. Focusing on the merged proportions of 1st and 3rd person subjects expressions (excluding NPs), the productions rates of overt pronouns were found to be higher at the intermediate level. Advanced learners, in contrast, produced more null subjects. Crucially, when the grammatical persons were separately considered, it was found that overt pronouns were expressed more often with 3rd than with 1st person verb forms. This confirms the 'selectiveness of deficits' claim of Lozano (2009b). The authors hypothesized that the increased proportion of overt pronouns in the lower proficiency may be due to the input received in classroom, where instructors may produce higher rates of overt pronouns than normal (Dracos, 2010). Regarding frequency effects, the authors concluded that for less proficient learners pronouns are expressed more often with frequent than with infrequent verbs.

---

one chose 6. According to their self-ratings, learners may be classified as of intermediate to lower-advanced proficiency (Quesada, 2015:226).

Gudmestad, House, & Geeslin (2013) was the first (and only to date) variationist study on L2 Spanish that focused exclusively on 3rd person subjects[98]. The authors used a Bayesian multinomial probit model to examine the production of native speakers and very advanced English-speaking learners of Spanish. Crucially, the sophisticated statistical method employed in this study allows to simultaneously consider the influence of multiple linguistic factors in the production of anaphoric subjects[99] (including lexical noun phrases, demonstrative, interrogative and indefinite pronouns). The results indicated that several of these factors are important, although only certain parameters of these factors predicted use. The subtle differences between L2ers and native speakers were found to specifically concern lexical noun phrases and some pronominal forms (demonstrative, interrogative and indefinite) but, crucially, not the traditionally-studied alternation between null and overt pronouns. Overall, in line with the results of the previous studies (Geeslin & Gudmestad, 2008, 2011; Gudmestad & Geeslin, 2010), learners and native speakers were found to be sensitive to the same constraints, although not to the same degree.

Recently, Geeslin, Linford, & Fafulas (2015) conducted the first variationist study on the L2 acquisition of anaphoric subjects that considers developmental issues. In contrast to previous studies, the authors examined English-speaking learners of six different proficiency levels[100] (ranging from beginners to graduate students in university Spanish courses). The responses of L2ers in a written contextualized task (WCT) were contrasted to those of native speakers. It should be noted that 1st and 3rd person were merged and the full range of forms was not examined in this study: the participants were asked to fill in

---

[98] For the need of separately examining 3rd person anaphoric subjects due to the lack of uniformity in the pronominal paradigm see section 2.3. The idea that this disunity may contaminate the results also appears in several recent variationist studies (Geeslin & Gudmestad, 2016; Shin & Cairns, 2012; Shin & Otheguy, 2009; Travis & Cacoullos, 2012).

[99] The traditional variationist methodology strictly requires the dependent variable to be binary (null vs overt). Additionally, the importance of each linguistic factor may only be assessed individually (Gudmestad, House, & Geeslin, 2013:376).

[100] Gudmestad & Geeslin (2010:273) early on recognized that the analysis of only one group of learners does not permit to make observations about development.

the missing subjects in the WCT with either null or overt pronouns (NPs were not among the options). The overall distribution of null and overt pronouns revealed that learners exhibit an inverted U-shaped developmental pattern, insofar as they start with low rates of overt pronouns (similar to those of the native speakers), gradually increasing until intermediate proficiency levels (3rd year) and gradually decreasing again until they reach native-like rates at upper-advanced level. Significant differences were found between the intermediate learners and the native speakers whereas the upper-advanced group performed native-like. The authors argued against the possibility of cross-linguistic influence from English, given that the performance of the lowest proficiency L2ers (1st grade) regarding overt pronoun production rates was very close to that of the native group whereas other more advanced L2 groups (2nd, 3rd and 4th grade) diverged[101].

3.3.2 Pragmatic approaches on the acquisition of Spanish anaphoric subjects

Two discourse-oriented studies that focus on the acquisition of anaphoric subject in Spanish L2 will be reviewed in this section (Blackwell & Quesada, 2012; Saunders, 1999). In line with the variationist works examined in the previous section, the authors of these studies focus on production data and depart from the assumption that the use of anaphoric subjects in discourse is constrained by multiple factors. However, in contrast to the probabilistic nature of the variationist approaches, the aforementioned authors examine the L2 acquisition of anaphora in pragmatic terms. Drawing heavily on the notions of saliency/topicality/givenness/activation and the information-structure constraints proposed under the traditional theoretical framework of discourse anaphora (see Chapter 2), pragmatic approaches seek to determine to what extent L2 learners and native speakers are affected by these constraints.

Saunders (1999) used oral production data (picture-based narrations) to contrast native speakers and English-speaking learners of Spanish at five proficiency levels[102] (from

[101] Other variationist studies have examined Spanish-English bilinguals in attrition contexts (native Spanish speakers living in EEUU) and have concluded that the contact with English causes an increase in the use of overt pronouns (Otheguy & Zentella, 2012; Otheguy, Zentella, & Livert, 2007).

[102] It was not possible to have access to an original copy of this dissertation. The overview performed here is based on the corresponding reviews provided in Quesada (2015:153-168) and Quesada & Blackwell (2009:117).

beginners to upper-advanced). The author proposed that the selection of anaphoric subjects in narrative discourse obeys to the rules of a universal hierarchy "based on the amount of information that the speaker/writer assumes the listener/reader already to possess" (p.51). The results indicated an overuse of noun phrases for the beginners and lower-intermediate learner groups. Overt pronouns were rarely used at these low-proficiency levels. This was interpreted as evidence against negative transfer from English (which would foster instead the overuse of overt pronouns). However, at the upper-intermediate and lower-advanced level, learners were found to overuse overt pronouns "using a system that more closely resembles that of English (the learners' L1)" (p.127). This finding confirms formal/generative studies who tested English-speaking learners of the same proficiency level (Lozano, 2009b, 2016; Montrul & Rodríguez Louro, 2006; *inter alia*). Only very advanced learners performed native-like (a finding that is in line with Rothman, 2009). The author concluded that "both native speakers and learners are constrained by the predictions of the anaphoric hierarchy, albeit in different ways until advanced levels" (Quesada, 2015:168).

Very similarly, Blackwell & Quesada (2012) examined native speakers and English-speaking learners of Spanish at three proficiency levels (beginner, intermediate and advanced). The authors tested the predictions of a revised adapted-to-Spanish version of the Givenness Hierarchy[103] (Gundel, Hedberg, & Zacharski, 1993) regarding the use of 3rd person anaphoric subjects in oral narrative data (film-retell task). Regarding the 'in focus' status, where null subjects are almost categorically expected, all L2 groups were found to overuse overt pronouns and noun phrases. Interestingly, native speakers also employed some redundant overt subjects in this condition, although to a significantly lesser degree. The authors illustrated through some examples that this was due to stylistic reasons ('emphasis'). In the 'activated and recoverable' status, where null subjects are also generally expected, beginners overused noun phrases and intermediates overused overt pronouns, whereas the advanced group was native-like. The authors interpreted these results in terms of potential cross-linguistic influence from English: "Learners must

[103] See 2.4.2.3 for the theoretical tenets of this model as well as its predictions regarding the correspondence between referential forms and cognitive statuses. As already discussed, the application of this model to direct measurements of 'givenness' in real discourse production entails some methodological problems. More specifically, as the authors also acknowledge, "it is impossible to directly assess the 'cognitive' status of referents in the mind of speakers" (Blackwell & Quesada, 2012:142).

learn that they do not need an overt subject when the referent is activated and the context makes it clear about whom we are speaking" (p.156). Regarding the 'activated' cognitive status, where the selection of an overt pronoun is predicted, native speakers were found to choose ambiguous null subjects in 12% of the instances. This unexpected referential choice that may lead to a breakdown in communication only concerned the production of the control group (no significant differences were found, however, between native speakers and learners). The same phenomenon was observed for the 'familiar' cognitive status, whereas the limited number of tokens in the remaining statuses did not permit conclusive statements to be made. The authors concluded that "subject expression for both NSs and learners at all levels is constrained by the cognitive status of discourse entities" (p.161). Overall, they argued that native speakers are more likely to use less specific forms whereas learners follow an acquisitional process of replacing more specific forms with more minimal ones as proficiency grows.

3.3.3 Summary of results of the discourse-oriented literature

The discourse-oriented studies reviewed in this section consensually confirm the existence of some universal factors of syntactic and pragmatic nature that govern the use of anaphoric subject expressions in both L1 and L2 Spanish. More specifically, English-speaking learners of L2 Spanish are reported to be sensitive to the same constraints that account for the referential choices of the native Spanish speakers, though not always to the same degree. It should be noted that this is not surprising, given that the theoretical literature on discourse anaphora has traditionally alleged the universality of these constraints (see Chapter 2). In other words, there is nothing particularly novel regarding the main finding of variationist and pragmatic approaches, insofar as the referential choices of both L2 learners and native speakers are expected to universally reflect the information status of the referent (see section 2.5). However, due to the different inventories of each language, it is reasonable to assume that the same information status is encoded differently in different languages (e.g. the highest information status in Spanish is encoded with null subjects whereas in English it is encoded with overt pronominals).

It should also be noted that, by focusing primarily on the aspects where learners and native speakers converge, variationist and pragmatic approaches have not been particularly concerned with the issues of 'optionality' and 'fossilization' as these have been addressed under formal/generative approaches. They do, however, provide some evidence in favour

of the idea that learners of L2 Spanish may have problems acquiring the discursive properties of anaphoric subjects. Early variationist studies demonstrated that advanced L2ers produce overall more null subjects, although the appropriateness of this production was not addressed. In more recent studies, where grammatical persons were separately examined, it was found that intermediate-proficiency learners overproduce 3rd person overt pronouns. This is fully in line with the primary consensual finding of the 'syntax-discourse interface' literature which has also been confirmed by studies conducted under pragmatic approaches. Additionally, the extensive review of the literature drives us to the conclusion that the non-target performance of learners does not equally affect all grammatical persons and anaphoric subject forms. On one hand, this crucial observation highlights the need to separately examine 1st, 2nd and 3rd person anaphors. On the other hand, the consideration of the entire range of anaphoric forms is equally essential. In sum, there is one important finding in which the bulk of formal and discourse-oriented approaches coincide, namely: intermediate-proficiency English-speaking learners of L2 Spanish overproduce 3rd person overt subjects. Bearing in mind the summary of findings from both formal and discourse-oriented approaches, the research questions and hypotheses of this thesis will be presented in the next chapter.

CHAPTER 4

4 RESEARCH QUESTIONS AND HYPOTHESES

This chapter deals with the specific research questions addressed in this thesis and the corresponding hypotheses that will be tested. Based on the theoretical background on discourse anaphora presented in Chapter 2 and the literature review on the L2 acquisition of anaphoric subjects in Chapter 3, we are now in a position to examine the following research questions and hypotheses:

Research question 1: Regarding the production of 3rd person anaphoric subjects in Spanish L1, what factors may account for the referential choices of the native Spanish speakers?

Hypothesis I: The referential choices of the native Spanish speakers will be constrained by the factors proposed in the theoretical literature on discourse anaphora. More specifically, we expect the information status of 3rd person referents in subject position to be reflected in the referential choices made by native speakers of Spanish according to the predictions made in the theoretical models on discourse anaphora (Ariel, 1990; Givón, 1983; Gundel, Hedberg, & Zacharski, 1993; Kibrik, 2011) and the complementary psycholinguistic, computational and LCR approaches (Arnold, 1998; Gudmestad, House, & Geeslin, 2013; Lozano, 2016; Mitkov, 2002; *inter alia)*. Less specific referential expressions (null subjects) are expected to be produced for more topical/activated/accessible/salient referents (i.e. higher information status) and more specific expressions (overt subjects) are expected to be produced for less topical/activated/accessible/salient referents (i.e. lower information status). The information status of the referent at each moment in discourse will be determined from the interaction of the two main factors that have been proposed in the theoretical literature, namely the referential distance ('Distance') and the interference from other referents ('PRI'). Additionally, several other factors that have been proposed more recently are also expected to contribute in the topicality/activation/accessibility/saliency of the referent, namely: Switch Reference, Clause Type, Priming, Antecedent Syntactic Function, Protagonisthood, New Paragraph and Shared Knowledge (see Chapter 5 for more details on each of these factors).

Research question 2: (**a**) Do learners of Spanish from both L1 backgrounds (English and Greek) show deficits in the production of 3rd person anaphoric subjects? (**b**) May the properties of 3rd person anaphoric subjects in Spanish L2 be eventually acquired?

Hypothesis II: (**a**) Learners of both L1 backgrounds (English and Greek) will exhibit non-target performance (deficits). More specifically, we expect both English-speaking and Greek-speaking learners of Spanish to show deficits in the production of 3rd person anaphoric subjects, though not necessarily to the same degree. English-speaking learners are expected to perform in a non-target fashion, in line with the findings of previous studies on the L2 acquisition of anaphoric subjects in Spanish (Lozano, 2009b, 2016; Montrul & Rodríguez Louro, 2006; Pérez-Leroux & Glass, 1997, 1999). Similarly, Greek-speaking learners are also expected to exhibit non-target performance, as previous literature has demonstrated (Lozano, forthcoming; Margaza & Bel, 2006). (**b**) Deficits will persist even at the upper-advanced levels of proficiency. According to the IH, deficits will be observed even in the production of the more advanced learners, irrespective of their L1 background (Sorace, 2011, 2016; Sorace & Filiaci, 2006; Sorace, Serratrice, Filiaci, & Baldo, 2009). More specifically, and in line with the findings of previous literature (Lozano, forthcoming, 2016), we expect both the upper-advanced English-speaking learners and the upper-advanced Greek-speaking learners to perform in a target-deviant fashion.

Research question 3: If non-target L2 performance is observed, does it concern overexplicitness, underexplicitness or both?

Hypothesis III: (**a**) Non-target L2 performance will primarily concern the overuse of redundant overt anaphoric subjects. More specifically, English-speaking and Greek-speaking learners of Spanish are expected to produce overexplicit 3rd person anaphoric subjects to a greater extent than native Spanish speakers. In line with previous research, overexplicitness is expected to be particularly evident in the interlanguage of intermediate proficiency learners of L2 Spanish (Al-Kasey & Pérez-Leroux, 1998; Lozano, forthcoming, 2009b; Margaza & Bel, 2006; Montrul & Rodríguez Louro, 2006; Pérez-Leroux & Glass, 1997). (**b**) Deficits will also concern, to a lesser degree, the overuse of ambiguous null anaphoric subjects. Against the UDH (Sorace, 2004, 2006a), and in line with the results of other previous studies (Lozano, 2009b; Montrul & Rodríguez Louro, 2006), learners are expected to overuse ambiguous null subjects to a greater extent than native Spanish speakers. In sum, deficits are expected to be in both directions: redundancy

and ambiguity, although learners are expected to be more redundant than ambiguous, in line with the PPVH (Lozano, 2016).

Research question 4: Is L2 performance affected by proficiency level?

Hypothesis IV: Overall, more proficient learners will perform in a more native-like fashion than less proficient learners. Learners from both L1 backgrounds are expected to perform better as proficiency level grows. In line with the results of previous literature (K. L. Geeslin & Gudmestad, 2016; Lozano, 2009b; Margaza & Bel, 2006; Montrul & Rodríguez Louro, 2006), intermediate proficiency groups are expected to show more deficits than advanced groups, whereas advanced groups are expected to show more deficits than upper-advanced groups. In other words, the upper-advanced groups are expected to exhibit the more native-like performance between the three proficiency levels, in contrast to the intermediate groups that are expected to diverge more from the native control group than the higher proficiency groups, irrespectively of L1 background.

Research question 5: Is the L1 a facilitating factor in the acquisition of anaphoric subjects in Spanish L2?

Hypothesis V: Overall, Greek-speaking learners will perform in a more native-like fashion than English-speaking learners, due to the facilitating factor of their L1. In line with similar previous research (Bel, Sagarra, Comínguez, & García-Alcaraz, 2016; Lozano, 2002b, 2002c, 2008b, White, 1985, 1986), the learners whose L1 shares the same parametric options with the target language are expected to take advantage of this similarity. On the other hand, the learners whose L1 differs from the target language with respect to anaphora are expected to transfer the patterns of anaphoric distribution from their native language. Given that Greek, in contrast to English, is a pro-drop language such as Spanish, Greek-speaking learners are expected to perform more native-like than English-speaking learners at all proficiency levels. More specifically, intermediate Greek-speaking learners are expected to show less deficits than their English-speaking counterparts, advanced Greek-speaking learners are expected to show less deficits than their English-speaking counterparts and upper-advanced Greek-speaking learners are expected to show less deficits than their English-speaking counterparts.

Research question 6: If non-target L2 performance is observed, what factors may account for it?

Hypothesis VI: The non-target performance of L2 learners will be better accounted for in terms of an interaction of multiple factors. Given that deficits are expected (to some

degree) irrespectively of the source language of learners (Hypothesis II) and that negative cross-linguistic is also expected to affect the performance of English-speaking groups (Hypothesis V), no single factor will satisfactory account for the non-target performance of all L2 groups together. Instead, in line with the claims made in some recent studies (Ryan, 2015; Sorace, 2011) we expect the interaction of multiple factors to provide a more holistic explanation of non-target L2 performance. More specifically, other factors that have been proposed in previous literature and are expected to be potentially relevant in the acquisition of anaphoric subjects are: the unnatural input received in instructional contexts (Pladevall-Ballester, 2013; Rothman, 2009), influence from a previously learnt L2 (Polio, 1995), universal pragmatic principles (Lozano, 2016) and processing difficulties (Sorace, 2016).

CHAPTER 5

5 METHOD

The methodological approach of the present study is described in this chapter which is comprised of six sections. The first section focuses on the characteristics of the CEDEL2 corpus, which has served as the source where all the data originated from. In the second section, a full account regarding the participants and the selection of texts for this study is provided. This is followed by a description of the software that was used for the annotation and the analysis of the data. In the fourth and fifth section an extensive description of the annotation scheme and the categories of the tagset is provided. To conclude, the dataset that resulted from the implementation of the tagset in the selected texts is presented in the last section.

5.1 The corpus: CEDEL2

The empirical database of this study consists of a selection of texts from the CEDEL2 corpus[104]. CEDEL2 is a Spanish written corpus designed and collected online by Cristóbal Lozano (Lozano, 2009a; Lozano & Mendikoetxea, 2013). The collection of the data started at 2006 and is an ongoing process carried out through a designated webpage[105]. The database mainly consists of an L1 English – L2 Spanish learner corpus and a native Spanish corpus. The explicit principles of Sinclair (2005) were applied to the design of the CEDEL2 learner corpus, as described in Lozano & Mendikoetxea (2013). More specifically, all participants have to fill in three online forms which are aimed to collect the following data (form samples are provided in Figure 48, Figure 49 and Figure 50 in the Appendix):

i. Social and learning background information: the first online form contains personal details about the participants (age, sex, email etc.) and, most importantly, linguistic details regarding their learning background (L1, length of instruction in Spanish, length of stay in Spanish-speaking countries, other foreign languages

[104]CEDEL2 stands for "Corpus Escrito Del Español L2" ("L2 Spanish Written Corpus").

[105] Data from English-speaking learners of Spanish are being collected online at this webpage: goo.gl/0s8hZV. Data from native speakers of Spanish are being collected online here: goo.gl/O8iCjN.

etc.). Additionally, the form contains a personal self-rating of proficiency in Spanish language, which may serve as a subjective measurement of linguistic competence.

ii. A standardized placement test (University of Wisconsin, 1998): the objective measurement of proficiency by the use of a standardized placement test guarantees the valid comparability of participants in terms of linguistic competence. This may be further combined with the learning background information and the self-evaluation data which, as we have seen, are collected in the previous form. Together, they constitute a robust methodological tool which may serve to determine the proficiency level of each participant in a highly accurate manner.

iii. A composition in Spanish (12 topics to choose from): the participants may choose among different topics which are aimed to cover a wide range of discourse genres (argumentative, narrative, expository and descriptive). Thus, the produced discourse may serve for the study of the learners' interlanguage regarding a variety of linguistic structures and phenomena.

Very recently, for the aims of this particular study, a complementary L1 Greek – L2 Spanish learner corpus was compiled and added to the CEDEL2 database. The collection of the Greek learners' data started at 2015 and is being performed through a designated online software[106]. Exact copies of the online forms that have been used for the collection of the Spanish L1 and English L1 – Spanish L2 data were designed and used for the compilation of the Greek L1 – Spanish L2 subcorpus. Therefore, the three data collections that comprise the actual version of the CEDEL2 corpus are fully comparable, since they were designed and collected with the same exact criteria. The data are, thus, ideal for comparative purposes and, more specifically, for the contrastive analysis that will be performed in this study.

[106] Data from Greek-speaking learners of Spanish are being collected online at the following webpage: https://test.ugr.es/limesurvey/index.php/282758/lang-es

Note here that all the data of the CEDEL2 corpus are being exclusively collected online. Their collection is being followed by an extensive data cleansing procedure and, successively, they are uploaded to a publicly available interface on the web[107]. The online collection is, in addition, fully browsable through a sophisticated search engine which allows several user-defined filters to be applied to the data. Through a straightforward selection of the relevant (according to research interests) variables, any collection of texts can be viewed online and/or downloaded (see Figure 51 in the Appendix for a graphical illustration of the CEDEL2 interface). Recall here that, through this procedure, a complete access to all the data of this study is provided. The participants' biodata and all the texts that were analysed are fully accessible with just a few mouse clicks. Therefore, the present study is one of the first that allows full open access to all off its data.

Regarding the size of the corpus, up to now it contains a total of more than 800.000 words (Spanish natives: 220.000 words, English-speaking learners: 510.000 words, Greek-speaking learners: 80.000 words) originating from more than 2.500 participants (800 Spanish natives, 1600 English-speaking learners, 173 Greek-speaking learners). For the aims of this study, data from the three subcorpora were selected (Spanish natives, English-speaking learners of Spanish, Greek-speaking learners of Spanish) according to specific criteria which will be made explicit in the following section.

5.2 The participants and the sample

Given the focus of this study on 3rd person anaphoric subjects, preference was given to texts that would be more likely to contain numerous animate referents and anaphoric relations (see also section 2.3). Therefore, only 'narrative' compositions ("Summarize a film that you have watched recently")[108] along with 'expository' ones ("Write about a famous person") were extracted from the corpus. However, during the analysis, the distinction between narrative and expository texts was found to be less clear-cut than expected. Most of the expository essays were found to contain narrative passages as well.

---

[107] http://cedel2.learnercorpora.com

[108] Only film retellings that focus on the film summary were selected. In contrast, those that simply criticize the film from an artistic point of view without narrating the story were discarded. The reason is that the latter category is less likely to contain 3rd person animate referents and anaphoric relations.

Therefore, it might be more accurate to describe the discourse genre under study as narrative, with some expository passages.

Following the aforementioned criteria, texts were extracted from seven groups of participants (the authors of the texts) divided according to their proficiency level in Spanish language. The final sample consists of 20 Spanish native speakers[109] and 72 Spanish learners[110] distributed into 6 groups according to their L1 and proficiency level. A summary of the participants' features is provided in Table 2:

Group	N	Mean age	Mean placement test score (%)	Mean self-evaluation score (1-6)[111]	Mean length of instruction (years)
Natives	20	30	n/a	n/a	n/a
English1	12	26	61%	2.6	4
English2	12	28	85%	3.9	6
English3	12	39	97%	5.2	10
Greek1	12	32	60%	1.5	1
Greek2	12	31	82%	3.8	3
Greek3	12	33	98%	5.3	5

Table 2. Summary of the participants' biodata and proficiency-related features

[109] Caribbean Spanish speakers were excluded since an overuse of overt subject pronouns has been reported in these varieties (Otheguy & Zentella, 2012; Toribio, 2000).

[110] The number of 12 participants in each learner group (and 20 in the control group) is relatively high when contrasted with the participant pools of other similar studies, e.g. 10 learners per proficiency group are examined in several recent production-oriented SLA studies on anaphoric subjects (Abreu, 2009; Blackwell & Quesada, 2012; Chini, 2009; Hendriks, 2003; Lozano, 2009b, 2016; Margaza & Bel, 2006; Ryan, 2015) whereas even less than 10 are examined in some studies (García-Alcaraz & Bel, 2011; Linford & Shin, 2013). Note, additionally, that none of the aforementioned studies examines more than one or two proficiency groups.

[111] The 6 degrees of the self-evaluation scale roughly correspond to the 6 levels of the CEFR (Council of Europe, 2001).

Learners of both L1s (English and Greek) were selected according to their score in the independent placement test. Intermediate learners (English1 and Greek1) scored between 49 and 70%, whereas advanced learners (English2 and Greek2) scored between 74 and 91% and upper-advanced learners (English3 and Greek3) scored more than 95%. A similar division has been applied to other previous CEDEL2 studies (Lozano, 2009b, 2016). Note here that although the criteria for dividing the participants into proficiency levels are objective (placement test scores), the labels regarding the proficiency of each group are subjectively assigned. One could argue, for example, that the intermediate learners should be labelled 'elementary' and the upper-advanced might be labelled 'near-natives' instead. What is more relevant for the purposes of this study, however, is the clear-cut division between different proficiency groups and the comparability of same-proficiency groups with each other. As Ädel (2015:418) points out: "It is essential that the selected texts and the selected populations be maximally comparable. If the researcher is comparing 'apples and oranges', any findings showing differences or similarities between samples will be flawed". Additionally, the decision to base our analysis on group results does not preclude that individual variation may also be important for the phenomenon under study. However, this is the common practice in SLA production-oriented research on the acquisition of anaphoric subjects (Blackwell & Quesada, 2012; Geeslin & Gudmestad, 2008, 2011, 2016; Geeslin, Linford, & Fafulas, 2015; Gudmestad & Geeslin, 2010; Gudmestad, House, & Geeslin, 2013; Lozano, 2009b, 2016; Montrul & Rodríguez Louro, 2006; Quesada & Blackwell, 2009). Moreover, as Geeslin & Gudmestad (2011:22) point out, "this decision is justified in the literature and makes sense from a practical standpoint". The authors cite additionally some recent research (Bayley & Langman, 2004; Regan, 2004) which shows that "despite individual variation the constraints on group behaviour are similar to those that govern individual language use" (p.22).

The biodata of the participants of the native control group can be seen in Table 3:

ID	Sex	Age	Country of origin
16_3_Bv	female	16	México
17_3_TIQUI	female	17	México
19_2_NNR	female	19	España
21_3_CPV	female	21	España
21_3_PMA	female	21	España
21_3_TW	female	21	Argentina
24_3_JAO	male	24	España
25_3_mj	female	25	España
26_2_AVS	female	26	España
26_2_JGB	male	26	España
26_3_EMP	female	26	España
28_3_JF	male	28	España
30_3_SG	female	30	España
32_3_MDD	female	32	España
40_2_JJMP	male	40	España
40_3_BSN	female	40	España
43_2_ERS	female	43	España
44_2_ASJ	male	44	México
44_2_CMR	female	44	España
44_2_PAC	male	44	España
AVERAGE:		**30**	

Table 3. Native control group

The native control group is comprised of 20 native speakers of Spanish, originating from Spain (n=16), Mexico (n=3) and Argentina (n=1). The mean age of the participants of this group is 30 years. There are 14 females and 6 males in the group. Full details for each participant of this group are available online (http://cedel2.learnercorpora.com/natives/).

The biodata of the participants of the English1 group can be seen in Table 4:

ID	Sex	Age	Proficiency score	Self-evaluation score (1-6)	Years of instruction
22_18_5_2_KAC	female	18	51%	3.5	5
23_20_5_2_RW	female	20	53%	2	5
23_23_3_2_JP	male	23	53%	2	3
24_19_6_2_SH	female	19	58%	3.5	6
25_66_4_3_HR	female	66	58%	1.5	4
26_19_4_2_CM	female	19	60%	3.5	4
27_19_4_3_OM	female	19	63%	3.5	4
27_33_2_2_LF	female	33	63%	1.25	2
29_20_ _2_EMH	female	20	67%	3.5	?
29_22_2_2_ALK	female	22	67%	2	2
30_20_6_3_NJP	male	20	70%	3	6
30_29_5_2_TS	male	29	70%	3	5
AVERAGE:		**26**	**61**	**2.7**	**4**

Table 4. English1 group (intermediate proficiency)

The participants of the English1 group are native speakers of English who scored between 51 and 70% in the standardized placement test (mean score 61%). There are 9 female and 3 male participants (mean age 26 years). Their mean length of instruction is 4 years and the mean self-evaluation score of this group is 2.7. Full details for each participant of this group are available online (http://cedel2.learnercorpora.com/learners-english/).

The biodata of the participants of the English2 group can be seen in Table 5:

ID	Sex	Age	Proficiency score	Self-evaluation score (1-6)	Years of instruction
33_18_4_2_MAN	female	18	77%	3.75	4
34_27_7_2_NJF	female	27	79%	3.5	6.5
35_15_7_3_LMR	female	15	81%	4.5	7
36_18_6_2_AF	male	18	84%	4	6
36_21_6_3_JSB	female	21	84%	3.75	6
37_17_9_2_CJR	female	17	86%	4.25	9
37_18_7_3_EM	female	18	86%	4	7
37_22_2_3_DH	female	22	86%	4	2
37_74_5_3_RRA	male	74	86%	3.25	5
38_19_4_3_MTR	female	19	88%	3.75	4
38_57_7_2_jd	male	57	88%	4.25	7
39_30_4_2_CLR	female	30	91%	4	4
AVERAGE:		**28**	**85**	**3.9**	**6**

Table 5. English2 group (advanced proficiency)

The participants of the English2 group are native speakers of English who scored between 77 and 91% in the standardized placement test (mean score 85%). There are 9 female and 3 male participants (mean age 28 years). Their mean length of instruction is 6 years and the mean self-evaluation score of this group is 3.9. Full details for each participant of this group are available online (http://cedel2.learnercorpora.com/learners-english/).

The biodata of the participants of the English3 group can be seen in Table 6:

ID	Sex	Age	Proficiency score	Self-evaluation score (1-6)	Years of instruction
41_28_15_3_KDH	female	28	95%	5	15
41_30_8_3_JM	male	30	95%	5	8
41_19_5_3_AEM	female	19	95%	5	4.5
41_37_12_3_CJD	female	37	95%	5	12
41_57_10_3_SME	female	57	95%	5	10
42_20_8_3_JEL	female	20	98%	5.5	8
42_21_10_3_LBK	male	21	98%	6	10
42_40_4_2_CMJ	female	40	98%	5	4
42_47_29_3_TLS	male	47	98%	6	29
42_48_11_3_OPE	male	48	98%	5	11
42_51_6_2_LP	female	51	98%	5	6
42_66_5_3_LML	male	66	98%	5	5
AVERAGE:		**39**	**97**	**5.2**	**10**

Table 6. English3 group (upper-advanced proficiency)

The participants of the English3 group are native speakers of English who scored more than 95% in the standardized placement test (mean score 97%). There are 7 female and 5 male participants (mean age 39 years). Their mean length of instruction is 10 years and the mean self-evaluation score of this group is 5.2. Full details for each participant of this group are available online (http://cedel2.learnercorpora.com/learners-english/).

Several unpaired t-tests were performed in order to ensure that the three English-speaking groups differ from each other with respect to proficiency level. Regarding proficiency scores, there was a significant difference between the English1 (M=61.08, SD=6.67) and the English2 group (M=84.67, SD=3.98); t=10.51, p<.0001. There was also a significant difference between the English2 (M=84.67, SD=3.98) and the English3 group (M=96.75, SD=1.54); t=9.79, p<.0001. The same procedure was followed for the self-evaluation scores of the three groups. There was a significant difference between the English1 (M=2.68, SD=0.87) and the English2 group (M=3.91, SD=0.34); t=4.53, p<.001. There was also a significant difference between the English2 (M=3.91, SD=0.34) and the English3 group (M=5.2, SD=0.39); t=8.53, p<.0001. Finally, the length of instruction of

the participants from the three groups was also compared. No significant difference (marginally) was found between the English1 (M=4.18, SD=1.4) and the English2 group (M=5.62, SD=1.89); t=2.06, p=.0521. On the other hand there was a marginally significant difference between the English2 (M=5.62, SD=1.89) and the English3 group (M=10.2, SD=6.78); t=2.25, p=.0344. Overall, the results indicate that the three English-speaking group can be safely differentiated with respect to proficiency level.

Following the same order of presentation, we start with the biodata of the participants of the Greek1 group that can be seen in Table 7:

ID	Sex	Age	Proficiency score	Self-evaluation score (1-6)	Years of instruction
21_22_1_2_JUA	Female	22	49%	1	1
21_24_1_2_ELE	Female	24	49%	1	1
22_21_0_2_christos	Male	21	51%	1	?
22_37_0_2_OK	Female	37	51%	1	?
23_37_1_2_DOM	Female	37	53%	1	1
24_19_2_2_Gar	Female	19	56%	3	2
26_46_1_3_MAR	Female	46	60%	2	0.7
28_37_1_2_OK	Female	37	65%	2	1
29_37_1_2_EVT	Male	37	67%	2	1
30_25_1_2_Katerina	Female	25	70%	1	1
30_49_1_3_D.K	Female	49	70%	1	0.5
30_25_1_2_GEO	Male	25	70%	1	1
AVERAGE:		**32**	**59%**	**1.5**	**1**

Table 7. Greek1 group (intermediate proficiency)

The participants of the Greek1 group are native speakers of Greek who scored between 49 and 70% in the standardized placement test (mean score 59%). There are 7 female and 5 male participants (mean age 32 years). Their mean length of instruction is 1 year and the mean self-evaluation score of this group is 1.5. Full details for each participant of this group are available online (http://cedel2.learnercorpora.com/learners-greek/). It should

be noted that all but one participant of this group (24_19_2_2_Gar) have reported some knowledge of English. This is normal, given that 'English as a Foreign Language' is an obligatory course since the first years of elementary school in Greece.

The biodata of the participants of the Greek2 group can be seen in Table 8:

ID	Sex	Age	Proficiency score	Self-evaluation score	Years of instruction
32_34_1_2_VASO	Female	34	74%	4	1
32_23_4_2_ELA	Female	23	74%	3	4
32_40_1_3_KAL	Female	40	74%	3	1
33_32_3_3_APO	Female	32	77%	4	3
34_42_1_2_YOR	Male	42	79%	3	1
35_19_4_3_SS	Female	19	81%	4	4
36_22_4_2_ELE	Female	22	84%	3	4
37_24_3_3_NAT	Female	24	86%	5	3
37_38_2_3_DIM	Male	38	86%	4	1.5
38_21_5_3_IFI	Female	21	88%	4	5
39_47_4_2_DOM	Female	47	91%	5	4
39_26_2_2_MAR	Female	26	91%	3	2
AVERAGE:		**31**	**82%**	**3.75**	**3**

Table 8. Greek2 group (advanced proficiency)

The participants of the Greek2 group are native speakers of Greek who scored between 74 and 91% in the standardized placement test (mean score 82%). There are 10 female and 2 male participants in this group (mean age 31 years). Their mean length of instruction is 3 years and the mean self-evaluation score of this group is 3.75. Full details for each participant of this group are available online (http://cedel2.learnercorpora.com/learners-greek/). Same as with the Greek1 group, all but two participants of this group (34_42_1_2_YOR and 36_22_4_2_ELE) have reported some knowledge of English.

Finally, the biodata of the participants of the Greek3 group can be seen in Table 9:

ID	Sex	Age	Proficiency score	Self-evaluation score	Years of instruction
41_32_3_2_TAL	Female	32	95%	5	3
41_39_7_3_FOT	Female	39	95%	5	7
41_36_10_3_SOF	Female	36	95%	6	10
42_26_2_3_ART	Female	26	98%	5	2
42_23_1_3_FA	Female	23	98%	4	1
42_36_8_3_MAR	Female	36	98%	6	8
42_29_4_3_MAR	Female	29	98%	5	4
42_38_3_3_DIM	Male	38	98%	6	3
43_46_2_3_TIM	Female	46	100%	6	2
43_33_7_2_NA	Female	33	100%	6	7
43_22_1_2_ATH	Female	22	100%	5	1
43_39_9_3_MAN	Female	39	100%	6	9
AVERAGE:		**33**	**98%**	**5.4**	**5**

Table 9. Greek3 group (upper-advanced proficiency)

The participants of the Greek3 group are native speakers of Greek who scored more than 95% in the standardized placement test (mean score 98%). There are 11 female and 1 male participants (mean age 33 years). Their mean length of instruction is 5 years and the mean self-evaluation score of this group is 5.4. Full details for each participant of this group are available online (http://cedel2.learnercorpora.com/learners-greek/). Same as with the Greek1 and Greek2 groups, all but one participant of this group (43_22_1_2_ATH) have reported some knowledge of English.

In line with the procedure followed for the English-speaking groups, several unpaired t-tests were performed in order to ensure that the three Greek-speaking groups differ from each other with respect to proficiency level. Regarding proficiency scores, there was a significant difference between the Greek1 (M=59.25, SD=8.72) and the Greek2 group (M=82.08, SD=6.47); t=7.28, p<.0001. There was also a significant difference between the Greek2 (M=82.08, SD=6.47) and the Greek3 group (M=97.92, SD=1.98); t=8.1, p<.0001. The same procedure was followed for the self-evaluation scores of the three

groups. There was a significant difference between the Greek1 (M=1.42, SD=0.67) and the Greek2 group (M=3.75, SD=0.75); t=8.02, p<.0001. There was also a significant difference between the Greek2 (M=3.75, SD=0.75) and the Greek3 group (M=5.42, SD=0.67); t=5.73, p<.0001. Finally, the length of instruction for the participants from the three groups was also compared. A significant difference was found between the Greek1 (M=1.02, SD=0.38) and the Greek2 group (M=2.79, SD=1.43); t=3.77, p=.0012. On the other hand, no significant difference (marginally) was found between the Greek2 (M=2.79, SD=1.43) and the Greek3 group (M=4.75, SD=3.25); t=1.9, p=.0695. Overall, same as with the English-speaking learners, the results indicate that the three Greek-speaking groups differ with each other with respect to proficiency level.

Finally, the same procedure was followed in order to compare the different-L1 groups (English-speaking vs Greek-speaking) across the three proficiency levels (intermediate, advanced, upper-advanced). Regarding their proficiency scores in the independent placement test, no significant difference was found between the English1 (M=61.08, SD=6.67) and Greek1 group (M=59.25, SD=8.72); t=0.58, p=.5687. Similarly, no significant difference was found between the English2 (M=84.67, SD=3.98) and the Greek2 group (M=82.08, SD=6.47); t=1.18, p=.2517. Finally, no significant difference was found between the English3 (M=96.75, SD=1.54) and the Greek3 group (M=97.92, SD=1.98); t=1.61, p=.1213. The results indicate that the English-speaking and Greek-speaking groups can be safely compared in terms of proficiency scores.

Although our only objective criterion in the classification of participants to proficiency levels was their proficiency score, we sought to find out whether the other two measures (self-evaluation and length of instruction) are also comparable between different-L1 groups of the same proficiency. We found that, regarding self-evaluation scores, English1 group participants (M=2.68, SD=0.87) self-evaluate themselves significantly higher than their Greek-speaking counterparts (M=1.41, SD=0.66); t=4, p<.001. No significant difference was found between the English2 (M=3.91, SD=0.34) and the Greek 2 group (M=3.75, SD=0.75); t=0.69, p=.4929. Similarly, no significance difference was found between the English3 (M=5.2, SD=0.39) and the Greek3 group (M=5.41, SD=0.67); t=0.92, p=.3632. Finally, regarding length of instruction, there was a significant difference between the English1 (M=4.18, SD=1.4) and the Greek1 group (M=1.02, SD=0.38); t=6.88, p<.0001, between the English2 (M=5.62, SD=1.89) and the Greek2 group (M=2.79, SD=1.43); t=4.12, p<.001, and between the English3 (M=10.2, SD=6.78) and Greek3 group (M=4.75, SD=3.25); t=2.51, p=.0197. This striking result indicates that

English-speaking learners need roughly the double time of instruction than Greek-speaking learners in order to score equally high to the same proficiency test.

Overall, the data of the participants confirm that, within groups, higher proficiency scores correspond to more years of instruction and higher self-evaluation rates. This is true for learners of both L1s (English and Greek) and for all the proficiency levels of each group. However, two observations are in order. First, intermediate English learners self-evaluate themselves higher than their Greek counterparts (2.6 vs 1.5)[112]. Second, and most important, all groups of Greek learners, with less exposure to Spanish language than their English counterparts, achieve the same or even higher proficiency score. In other words, English-speaking learners need the double (or more) of years of instruction in Spanish than Greek-speaking learners in order to score equally high in the same proficiency test. The latter observation has two direct implications. On one side, it can be taken as a first indication that L1 may be a facilitating factor in SLA. We will come back to this in Chapter 6 during the analysis of the results. On the other side, it points out the need for objective measurements of proficiency in SLA empirical research. We will argue together with Lozano & Mendikoetxea (2013:71) that knowing the objectively-defined proficiency level of each learner in the corpus is essential. Granger (2012:9) further notes that "proficiency level is often assigned on the basis of external criteria (number of years of study), an imperfect measure that has been denounced by a number of researchers". In the same line, Tono (2003:801) gives us a striking example of this imperfect measurement by examining Japanese learners of English from the ICLE corpus. He concludes that, although they are considered equally proficient with other students on the basis of length of exposure "their proficiency levels are so markedly lower than those from other European countries that the inclusion of the Japanese data seems to skew the overall results". Myles (2015:316) points out that "being in the same year group at school is not always a sufficiently rigorous indication, and it is advisable to carry out independent measures of proficiency". Consequently, there seems to be a general consensus about the need of objective assessment of proficiency levels in SLA research (see also Carlsen,

[112] It is out of the scope of this study to examine why this might be. It should be noted, however, that we are dealing with a personal subjective evaluation which is simply indicative. Therefore, a margin of diversity is to be expected.

2012; Thomas, 1994). The methodology of this study and the data presented in this section are fully in line with this.

5.3 The software: UAM CorpusTool

The dataset described in the previous section was imported in UAM CorpusTool which is an XML-based text annotation software[113] (O'Donnell, 2009). UAM CorpusTool has many powerful features including: design of a custom annotation scheme, annotation of multiple texts and annotation at multiple levels (e.g. word, clause, whole document etc.). Moreover, UAM CorpusTool allows for sophisticated search queries and provides descriptive and contrastive statistics between datasets according to user-defined criteria. It performs chi-square tests automatically for each comparison and reports the χ^2 value and the significance level[114]. Finally, it should be noted out that UAM CorpusTool is a stand-off annotator. Following Sinclair (2005:6), the annotated segments are "stored separately from the plain text and merged when required in applications". This allows to create multiple annotation layers whereas the original text stays untouched. All of the above features were used in the current study.

5.4 The design of the annotation scheme

The design of the annotation scheme was done according to the focus of this study and the corresponding previous literature. Given that we are dealing with 3rd person anaphoric subject expressions in real discourse, the complexity of the phenomenon under study does not allow for any kind of automatic annotation. Furthermore, given the nature and quantity of factors that have been previously suggested to affect the distribution of anaphoric subjects, only a fine-grained annotation scheme would be appropriate. More specifically, following Lozano & Díaz-Negrillo (submitted), an Interlanguage Annotation (ILA) tagset was designed and implemented to the data. By definition, ILA is a manual, fine-grained and purpose-oriented annotation procedure which has been widely used in studies on reference and anaphora in L2 discourse (Abreu, 2009; Blackwell & Quesada,

[113] UAM CorpusTool can be freely downloaded here: http://www.corpustool.com/.

[114] UAM CorpusTool reports the χ^2 value but not the exact p value. The significance level is only expressed in percentage terms: weak significance (90%), medium significance (95%) or high significance (98%).

2012; Chini, 2005; Geeslin & Gudmestad, 2016; Gudmestad, House, & Geeslin, 2013; Lozano, 2009b, 2016; Torregrossa & Bongartz, forthcoming).

The design of the annotation scheme of this study was largely inspired by the annotation schemes in Lozano (2009b, 2016). Lozano's tagsets (for the latest and more comprehensive version see Figure 52 in the Appendix) served as the original source upon which most annotation categories of the present study were designed. Additionally, the theoretical studies on discourse anaphora and previous research on LCR (see Chapter 2 and Chapter 3) were also taken into account in the design of the tagset. For this purpose, the vast tradition of empirical research on anaphora in linguistics, psycholinguistics and computational linguistics was thoroughly consulted. The present annotation scheme is intended to be the most fine-grained and integrated up until now in the study of anaphoric subjects in L2 Spanish. The design and implementation of the tagset, far from being a straightforward process, included the following steps:

- An initial version of the actual tagset was designed on basis of the more recent and relevant studies for this investigation (Blackwell & Quesada, 2012; Gudmestad, House, & Geeslin, 2013; Lozano, 2009b, 2016).
- The first version was experimentally implemented to some texts in order to pilot the annotation scheme. Some categories were found to be inappropriate for the analysis of the phenomenon under study and new categories were added.
- Several tags were revised and modified through a heuristic process according to the actual research focus. Particular emphasis was given to the operationalizability of the tags.

During the procedure, the entire text was first being read at least twice before the beginning of the annotation. Notes were being taken during this whole come-and-go procedure and a manual with the changes was being kept. Additionally, a list of special cases was created during the process. It should be noted here that the emphasis given to the objective definition of the tags is the most important guarantee of consistency in the tagging process. As Geeslin & Gudmestad (2011:21) point out, "inter-rater reliability with this type of coding scheme is by nature quite high". An extensive description of the tagset and the literature sources related to each of its categories will be presented in the next section.

5.5 The tagset categories

Given the size and the complexity of the tagset, in this section we will examine its categories one by one (see Figure 53 in the Appendix for the full annotation scheme). For each feature, the relevant literature sources will be cited and real discourse examples from the corpus will be provided. The first section deals with the features of the anaphoric expression and the second section with the features of the antecedent.

5.5.1 Anaphor's features

The anaphoric subject expression was tagged for the following features: Form, Number, Gender, Animacy and Clause Type. Each feature will be separately examined in the following sections. In all the examples, the relevant anaphoric subject forms are in bold.

5.5.1.1 Subject form

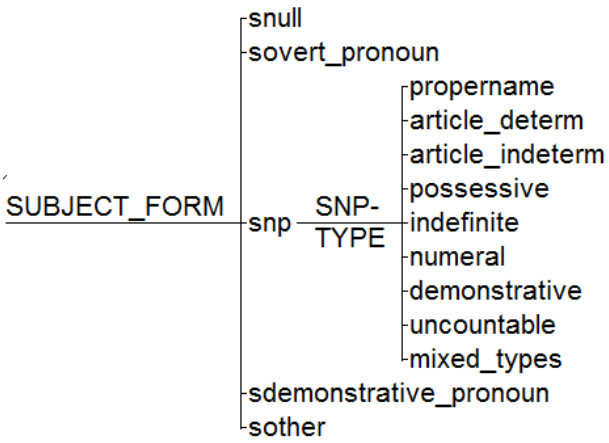


Figure 5. Subject form

The vast majority of previous studies on anaphoric subjects in Spanish typically consider only two forms: null and overt pronouns (see sections 3.2 and 3.3). Following some very recent studies (Blackwell & Quesada, 2012; Gudmestad et al., 2013; Lozano, 2009b, 2016) the entire range of 3rd person subject forms was considered here. A novelty of this study is that noun phrases were extensively annotated according to their type[115]. Furthermore, demonstrative pronouns were separately considered and all types of other

[115] Given that NPs have been only very recently considered in SLA studies on the acquisition of anaphoric subjects, the differences between each NP type could be further explored in future research.

subject forms (e.g. indefinite pronouns) were also tagged. Examples for each category and type are given below:

-snull

43) Él fue amable y **Ø** quiso ayudarlos (ENG22_18_5_2_KAC)
"He was kind and (**he**) wanted to help them"

-sovert_pronoun

44) **Él** nació el 10 de agosto de 1960 (GR21_22_1_2_JUA)
"**He** was born at 10 of August of 1960"

-snp/propername

45) **John Lehnon** murió asesinado a manos de un fan (ESP44_2_ASJ)
"**John Lehnon** was murdered by a fan"

-snp/article_determ

46) **Las monjas** viajaron muchos lugares (ENG33_18_4_2_MAN)
"**The nuns** travelled to a lot of places"

-snp/article_indeterm

47) Un día **una mujer** viene al colegio (ENG22_18_5_2_KAC)
"One day **a woman** comes to the college"

-snp/possessive

48) **Su esposo** ha empezado una clase (ENG41_19_5_3_AEM)
"**Her husband** has started a class"

-snp/indefinite

49) **Muchas mujeres** unieron su causa (ENG33_18_4_2_MAN)
"**Many women** joined her cause"

-snp/numeral

50) **Dos chicos** vean el accidente (ENG25_66_4_3_HR)
"**Two kids** see the accident"

-snp/demonstrative

51) **Este personaje** es de ascendencia china (ENG42_47_29_3_TLS)
"**This person** is of Chinese origin"

-snp/uncountable

52) **Cada padre** enseña a su hijo a luchar (GR42_23_1_3_FA)
"**Every father** teaches his sun to fight"

-snp/mixed_types[116]

53) **Ella y su hermano** son twins (ENG27_19_4_3_OM)
"**She and her brother** are twins"

-sdemonstrative_pronoun

54) **Este** aceptó inmediatamente (ENG38_57_7_2_jd)
"**He** accepted immediately"

-sother

55) **Juntos** tienen un hijo (GR32_21_1_2_christos)
"**Together they** have a son"

5.5.1.2 Number

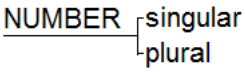


Figure 6. Number

Following Lozano (2009b), all subject expressions were tagged for number in order to allow for comparison between singular and plural forms. According to the author, English-speaking learners of Spanish show more deficits with 3rd person singular subjects than with plural ones. Examples of both categories are given below:

-singular

56) **Nadal** es muy famoso (GR30_25_1_2_Katerina)
"**Nadal** is very famous"

-plural

57) **Ellos** tienen un hijo nuevo (ENG23_20_5_2_RW)
"**They** have a new son"

5.5.1.3 Gender

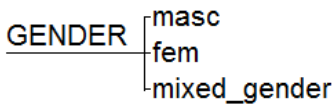


Figure 7. Gender

All subject expressions were tagged for grammatical gender. Descriptive categories such as gender, once thoroughly tagged, allow for specific search queries in order to identify particular cases of anaphora. Regarding the gender category, a novelty of this study is the

[116] Mixed refers to a combination of two or more anaphors of different types in the subject position.

'mixed' feature which refers to cases where a combined masculine and feminine referent occupy the subject position. The anaphor gender was also considered in Lozano (2016). See the following examples for each category:

-masc

58) **El hombre** murió (ESP19_2_NNR)
"**The man** died"

-fem

59) **Ella** es alta y delgada (ENG24_19_6_2_SH)
"**She** is tall and thin"

-mixed_gender

60) **Angelica y Brad** tenían otras esposos (ENG23_20_5_2_RW)
"**Angelica and Brad** had other husbands"

5.5.1.4 Animacy

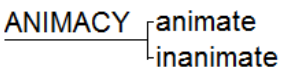


Figure 8. Animacy

The role of animacy in AR is crucial (Dahl & Fraurud, 1996; Fukumura & van Gompel, 2011). Particularly in Spanish, an overt subject pronoun can only refer to persons (Luján, 1999:1294). Moreover, Lozano (2009b) also considers the anaphor animacy and provides evidence that English learners of Spanish show deficits primarily in the use of referential animate subjects. In order to account for the effect of animacy, all subject expressions were tagged for this feature. Examples extracted from the corpus are given below:

-animate

61) **Cenicienta** está cantando (GR37_24_3_3_NAT)
"**Cindirella** is singing"

-inanimate

62) **La película** es sobre una poeta (ENG35_15_7_3_LMR)
"**The film** is about a poet"

5.5.1.5 Clause type

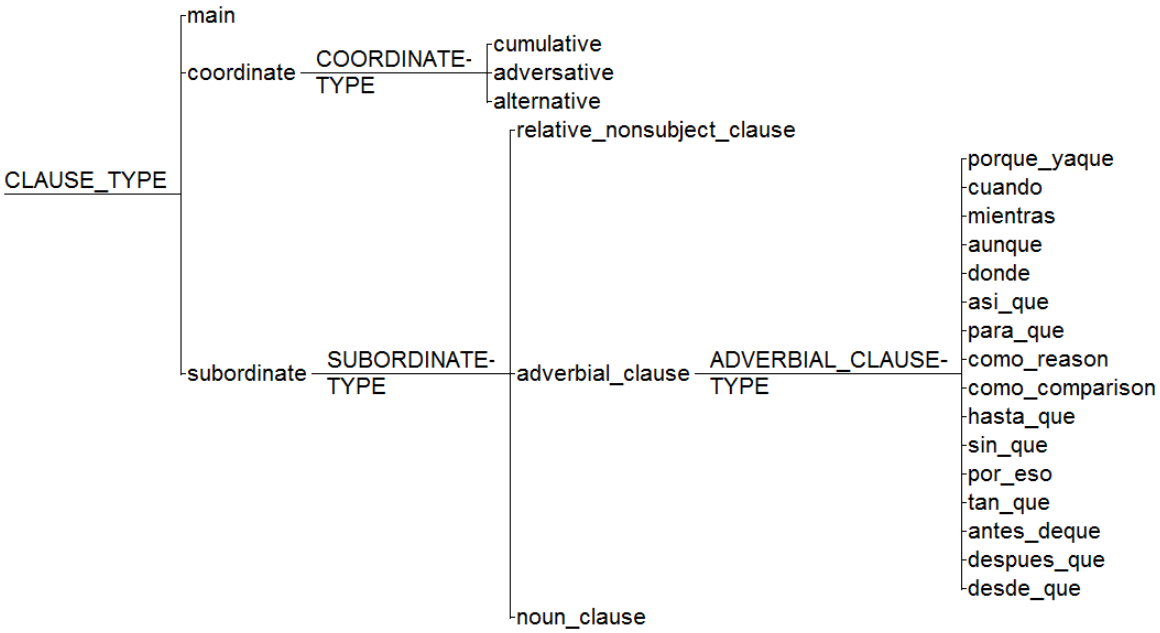


Figure 9. Clause type

Anaphora studies have traditionally considered the distinction between intra- and intersentential anaphora, according to clause type and order (main and subordinate clauses) whereas previous literature mostly focuses on the intrasentential type (main-subordinate clause order). On the other hand, clause type and order (main-subordinate or subordinate-main) has been considered as a potentially relevant factor in a number of recent studies (Bel & García-Alcaraz, 2015; García-Alcaraz, 2015; Liceras, de la Fuente, & Sanz, 2010; Margaza & Bel, 2006; Miltsakaki, 2007). Notice that the distinction typically concerns only main and subordinate clauses whereas, very recently, some studies have considered all three clause types (Otheguy & Zentella, 2012; Shin & Erker, 2015; Shin & Montes-Alcalá, 2014). Notice, however, that none of the aforementioned studies includes lexical NPs in the analysis. In the present study, the clause that contains the anaphor will be tagged for being either main, coordinate or subordinate. Additionally, all subtypes of coordinate and subordinate clauses will be considered. Crucially, subordinate adverbial clauses are classified according to their conjunction since its relevance for anaphora has been repeatedly pointed out in the literature[117] (Caramazza, Grober, Garvey, & Yates, 1977; Garvey & Caramazza, 1974; Hartshorne, Sudo, &

[117] This will allow to separately examine in the future certain clause types, e.g. causal clauses, where the implicit causality of the verb may affect referential choices (Goikoetxea, Pascual, & Acha, 2008).

Uruwashi, 2013; Miltsakaki, 2002; Stevenson, Knott, Oberlander, & McDonald, 2000). Examples extracted from the corpus for each clause category are given below:

-main

63) **Él** es mi actor favorito (ENG29_22_2_2_ALK)
"**He** is my favorite actor"

-coordinate/cumulative

64) Él es un futbolista portugués y **Ø** juega en el Real (GR30_25_1_2_GEO)
"He is a Portuguese football player and (**he**) plays for Real"

-coordinate/adversative

65) Es famosa pero **Ø** no es una actriz (ENG33_18_4_2_MAN)
"She is famous but (**she**) is not an actress"

-coordinate/alternative

66) Bolsos y maquillaje que ella quiere o **Ø** necesita (ENG29_20_ _2_EMH)
"Purses and make-up that she wants or (**she**) needs"

-subordinate/relative_nonsubject_clause

67) La persona con la cual **Adaline** se enamoró (GR38_21_5_3_IFI)
"The person with which **Adaline** fell in love"

-subordinate/adverbial_clause/porque_yaque

68) Me gusta mucho Banderas porque **Ø** es muy hermoso (GR21_22_1_2_JUA)
"I like Banderas a lot because (**he**) is very handsome"

-subordinate/adverbial_clause/cuando

69) Mas de 500 personas vio a Jesus cuando **Ø** estuvo en la tierra (ENG39_30_4_2_CLR)
"More than 500 persons saw Jesus when (**he**) came to the earth"

-subordinate/adverbial_clause/mientras

70) La niña necesita a su madre mientras **Ø** crece (GR32_40_1_3_KAL)
"The girl needs her mother while (**she**) grows up"

-subordinate/adverbial_clause/aunque

71) Decide conocerlo en persona aunque finalmente **Ø** se arrepiente (ESP28_3_JF)
"He decides to meet him in person although at the end (**he**) regrets it"

-subordinate/adverbial_clause/donde

72) A Emily la botan de la casa del señor donde **Ø** trabajaba (ENG41_37_12_3_CJD)

"Emily is fired from the house of the man where (**she**) worked"

-subordinate/adverbial_clause/asi_que

73) Danny necesito un cambio en su vida así **el** pide su novia a se casa (ENG30_20_6_3_NJP)

"Danny needed a change in his life so **he** asks his girlfriend to marry"

-subordinate/adverbial_clause/para_que

74) Llamando a los animales para que **Ø** la ayuden (ESP21_3_CPV)

"Calling the animals so that (**they**) help her"

-subordinate/adverbial_clause/como_reason

75) Como **Ø** no tiene independencia económica, no sabe qué hacer (GR43_46_2_3_TIM)

"Since **(he)** is not financially independent, (he) does not know what to do"

-subordinate/adverbial_clause/como_comparison

76) La madre decide cuidarla como **Ø** había cuidado a la tía (ENG42_20_8_3_JEL)

"The mother decides to look after her like (**she**) had looked after the aunt"

-subordinate/adverbial_clause/hasta_que

77) Todo parece perfecto hasta que **Marco** le dice a Verónica la verdad (ENG42_21_8_3_LBK)

"Everything looks perfect until **Marco** tells Verónica the truth"

-subordinate/adverbial_clause/por_eso

78) Él falta confianza y por eso **Ø** tiene que dominar su mujer (ENG41_19_5_3_AEM)

"He lacks confidence and that´s why (**he**) has to dominate his wife"

-subordinate/adverbial_clause/tan_que

79) Se sentía tan feliz que **Ø** no quisiera regresar a América (GR33_32_3_3_APO)

"He felt so happy that **(he)** did not want to return to America"

-subordinate/adverbial_clause/antes_deque

80) Rufus jugaba en muchos bares antes de que **Ø** ganó su contrato primero (ENG37_17_9_2_CJR)

"Rufus was playing in many bars before **(he**) won his first contract"

-subordinate/adverbial_clause/despues_que

81) Después de **el cazador** es muerto, Chigurh buscando la esposa del cazador (ENG25_66_4_3_HR)

"After **the hunter** is dead, Chigurh looking for the wife of the hunter"

-subordinate/adverbial_clause/desde_que

82) La Rosarina está con la "Pulga" desde que **Ø** eran chicos (GR32_34_1_2_VASO)

"Rosarina is with the "Flea" since (**they**) were children"

5.5.2 The antecedent

The antecedent was tagged for the following features: Switch Reference, Antecedent Form, Distance, Syntactic Function, PAS, Protagonisthood, New Paragraph, Active Referents and Shared Knowledge Constraints. Each feature will be treated separately in the following sections. In each example, the anaphoric subject of the clause under study is in bold. The antecedent and all other interacting referents are marked with subscript symbols.

5.5.2.1 Switch Reference

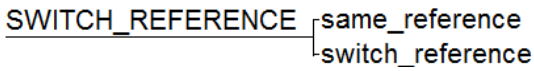


Figure 10. Switch Reference

Switch Reference accounts for the fact that a referring expression may co-refer with the syntactic subject of the previous clause (same-reference) or not (switch-reference). In other words, this tags examines whether the antecedent is in the subject position of the previous clause or not. This simple distinction has been found to drastically constrain the choice of referential expression in Spanish (Abreu, 2009; Bentivoglio, 1983; Cameron, 1994; Cameron & Flores-Ferrán, 2004; Flores-Ferrán, 2010; Geeslin & Gudmestad, 2016; Otheguy & Zentella, 2012; Shin & Cairns, 2012; Shin & Otheguy, 2009; Silva Corvalán, 1982, 1994). The general finding is that more explicit forms are expected in switch-reference than in same-reference contexts. It should be pointed out here that the same-reference and switch-reference categories approximately coincide with the topic-continuity and topic-shift terms that have been extensively employed in generative approaches on AR (Bel & García-Alcaraz, 2015; Jegerski, VanPatten, & Keating, 2011; Lozano, 2009b, 2016; Montrul, 2004a; Montrul & Rodríguez Louro, 2006; Papadopoulou, Peristeri, Plemenou, Marinis, & Tsimpli, 2015; Serratrice, 2007b; Sorace & Serratrice, 2009; Zulaica-Hernández, 2016). However, the lack of an objective definition of 'topic' makes them less appropriate for the annotation of anaphora in

production data (see section 2.4.1 for more details). In words of Slabakova, Kempchinsky, & Rothman (2012:323): "the discourse definitions of terms like 'topic' and 'focus' vary from one analysis to the next, and sometimes enter into direct contradictions". Switch Reference, on the other hand, is straightforwardly defined and operationalized. Natural discourse examples for the two Switch Reference types are given below:

-same_reference

83) Robert$_i$ había dejado su coche aparcado cerca del paso a nivel y **Ø$_i$** se había sentado sobre las vías (ESP44_2_CMR)

"Robert$_i$ had left his car parked near the grade crossing and **(he)$_i$** had sat on the rails"

-switch_reference

84) El$_i$ comienza a contarle$_j$ acerca de una muchacha$_k$ llamada Ali. **Ali$_k$** viene de una familia muy acomodada (ESP17_3_TIQUI)

"He$_i$ starts narrating her$_j$ about a girl$_k$ named Ali. **Ali$_k$** comes from a very well-off family"

5.5.2.2 Antecedent form

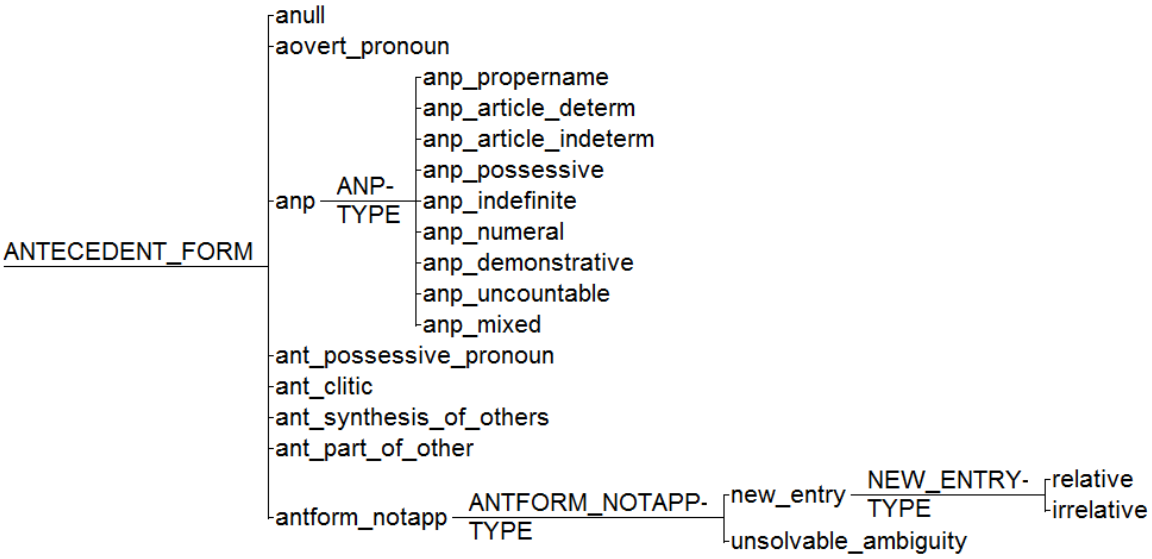


Figure 11. Antecedent form

Several studies in anaphora have found that the form of the antecedent affects the anaphoric choice. This phenomenon is generally known as priming (Cameron, 1994; Flores-Ferrán, 2002; Torres Cacoullos & Travis, 2014; Travis, 2005, 2007) or perseveration (Cameron & Flores-Ferrán, 2004; K. L. Geeslin & Gudmestad, 2016; Gudmestad, House, & Geeslin, 2013). In words of Travis (2007:102) "a preceding coreferential unexpressed (or implicit) subject tends to lead to a subsequent unexpressed subject and a preceding coreferential expressed (or explicit) subject tends to lead to a

subsequent expressed subject". Note, however, that most of the previous studies have focused on 1st person null and overt pronouns. In order to examine the effect of priming in 3rd person referents, the form of the antecedent was tagged in the same meticulous way and with the exact same categories that were used for Subject Form. Apart from the three basic forms (null, overt pronoun and NP) another two types of antecedent were identified and will be accounted for: possessive pronoun and clitic form. Additionally, regarding 3rd person plural anaphors, as Stirling & Huddleston (2010:1458) note: "Antecedents, namely so-called "split antecedents", can also be made up from two or more separate parts"(e.g. "John$_i$ loves Mary$_j$. **They$_{ij}$** are getting married"). Inversely, a 3rd person singular anaphor may refer to only a part of a group, as Soriano (1999:1216) has pointed out (e.g. "They$_{ij}$ are getting married, but **he$_i$** is not happy with it"). This is the first study that considers the form of the antecedent in such a fine-grained way. Finally, it should be pointed out that the Antecedent Form feature is not applicable to referents that due to unsolvable ambiguity are not possible to identify or appear for the first time. The new entries, however, are further tagged as being related to some previously mentioned referent or not. Examples extracted from the corpus for each category are given below:

-anull

85) Cuando Ø$_i$ nació, **ella$_i$** se llamó Chloe Wofford (ENG27_33_2_2_LF)
"When (she)$_i$ was born, **she$_i$** was named Chloe Wofford"

-aovert_pronoun

86) Él$_i$ aprendió que **él$_i$** tuvo otras hermanas y hermanos (ENG22_18_5_2_KAC)
"He$_i$ learned that **he$_i$** had other brothers and sisters"

-anp/anp_propername

87) Theodorakis$_i$ obtuvo una beca para estudiar en Paris donde **Ø$_i$** studio análisis musical (GR39_47_4_2_DOM)
"Theodorakis$_i$ got a scholarship for studying in Paris where **(he)$_i$** studied musical analysis"

-anp/anp_article_determ

88) La persona$_i$ que yo admiro mucho es famosa pero **Ø$_i$** no es una actriz (ENG33_18_4_2_MAN)
"The person$_i$ that I admire a lot is famous but **(she)$_i$** is not an actress"

-anp/anp_article_indeterm

89) Una plaga$_i$ pasa para Venecia y **Ø$_i$** mata a muchas personas (ENG42_21_8_3_LBK)

"A plague$_i$ hits Venice and **(it)$_i$** kills a lot of people"

-anp/anp_possessive

90) Su esposo$_i$ es un cantada también. **Él$_i$** es una parta del grupo (ENG26_19_4_2_CM)

"Her husband$_i$ is also a singer. **He$_i$** is a part of the group"

-anp/anp_indefinite

91) Normalmente, no voy a ver muchas películas$_i$ porque **Ø$_i$** son muy caros (ENG41_57_10_3_SME)

"Normally I don´t go to see many films$_i$ because **(they)$_i$** are very expensive"

-anp/anp_numeral

92) Los dos$_i$ usan armas ridículos porque **Ø$_i$** no tienen armas oficiales (ENG41_30_8_3_JM)

"The two$_i$ of them use ridiculous weapons because **(they)$_i$** don´t have official weapons"

-anp/anp_demonstrative

93) Ese hombre$_i$ famoso se llamo Alan Turing y **Ø$_i$** fue un matematico (GR39_26_2_2_MAR)

"This famous man$_i$ was named Alan Turing and **(he)$_i$** was a mathematician"

-anp/anp_mixed

94) El 12 de febrero de 2013 Penélope$_i$ y Javier$_j$ confirman que **Ø$_{ij}$** esperan su segundo hijo (GR32_23_4_2_ELA)

"At 12 of February of 2013 Penélope$_i$ and Javier$_j$ confirm that **(they)$_{ij}$** expect their second child"

-ant_possessive_pronoun

95) Su$_i$ madre$_j$ murio cuando **Cinderella$_i$** era muy joven (GR35_19_4_3_SS)

"Her$_i$ mother$_j$ died when **Cinderella$_i$** was very young"

-ant_clitic

96) Los soldados la$_i$ encuentran y al final **ella$_i$** se casa con el principe (GR37_24_3_3_NAT)

"The soldiers find her$_i$ and finally **she$_i$** gets married to the prince"

-ant_synthesis_of_others

97) El congresista$_i$ le acompaña al nino$_j$ y **Ø$_{ij}$** suben de nuevo al ascensor (ENG42_48_11_3_OPE)

"The congressman$_i$ accompanies the child$_j$ and **(they)$_{ij}$** get on the elevator again"

-ant_part_of_other

98) Cuando llegieron$_i$ a los Estados Unidos, **Barack$_i$** fue diez (ENG22_18_5_2_KAC)

"When they$_i$ arrived at the United States, **Barack$_i$** was ten"

-antform_notapp/new_entry/relative

99) Él tiene 54 años. **Su madre** era professora (GR21_22_1_2_JUA)
"He is 54 years old. **His mother** was a teacher"

-antform_notapp/new_entry/irrelative

100) Un día **una mujer** viene al colegio (ENG37_18_7_3_EM)
"One day **a woman** comes to the college"

-antform_notapp/unsolvable_ambiguity

101) El abogado$_i$ salva de la cárcel el policía$_j$, mientras el médico$_k$ opera al niño$_l$ - tomando un gran riesgo – y Ø$_k$ lo salva de la discapacidad. Un acontecimiento muy trágico da la última gota que colma el vaso. **Sus$_?$ hijos** matan con puñetazos y patadas una mujer sin hogar (GR43_39_9_3_MAN)
"The lawyer$_i$ saves the policeman$_j$ from prison, while the doctor$_k$ operates the child$_l$ – taking a big risk – and he$_k$ saves it$_l$ from disability. A very tragic event is the last straw. **His$_?$/Their$_?$ children** kill with punches and kicks a homeless woman"

5.5.2.3 Antecedent distance

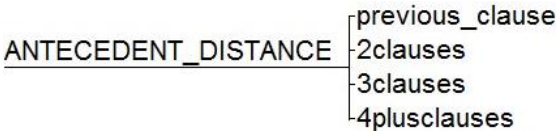


Figure 12. Antecedent distance

Since the seminal study of Givón (1983), the distance between two subsequent mentions of a referent has been consistently found to be one to the most relevant factors affecting the choice of referential form. In the literature, this phenomenon is widely known as Antecedent Distance or Recency (Abreu, 2009; Ariel, 1988; Arnold, 1998; Dumont, 2006; García-Alcaraz, 2015; Grüning & Kibrik, 2005; Gudmestad, House, & Geeslin, 2013; Kibrik, 1996, 2001; Lozano, 2016; Torregrossa, Bongartz, & Tsimpli, 2015). The general idea concerning Recency could be resumed to this: the longer the distance to the antecedent, the more explicit a referential form needs to be. In order to account for this, distance was measured linearly as the number of clauses that intervene between the anaphor and the antecedent. The traditional definition of clause was followed here: a group of words which contains a subject and a finite verb. According to Mitkov (2002:18): "empirical evidence suggests that the distance between a pronominal anaphor and its antecedent in most cases does not exceed 2–3 sentences". Therefore, all antecedents that were found to be more than four clauses away from the anaphor were

analyzed under the same tag. Examples for each category of Antecedent Distance are given below:

-previous_clause

102) Jessica$_i$ es una cantante bien y una actriz. **Jessica$_i$** teine pelo rubio y ojos azules (ENG24_19_6_2_SH)

"Jessica$_i$ is a good singer and actor. **Jessica$_i$** has blond hair and blue eyes"

-2clauses

103) La mujer$_i$ refinada del médico$_j$ desprecia la bella pero superficial mujer$_k$ del abogado$_l$. El médico$_j$ tiene un hijo$_m$ de dieciséis años. **El_abogado$_l$** tiene una hija$_n$ de quince años (GR43_39_9_3_MAN)

"The refined wife$_i$ of the doctor$_j$ looks down on the pretty but superficial wife$_k$ of the lawyer$_l$. The doctor$_j$ has a sixteen years old son$_m$. **The lawyer$_l$** has a daughter$_n$ who is fifteen years old"

-3clauses

104) El$_i$ intenta traerle$_j$ una alegría porque ella$_j$ sufre de una enfermedad que Ø$_j$ no recuerda nada de su$_j$ pasado ni aun a su$_j$ propia familia. **El$_i$** comienza a contarle$_j$ acerca de una muchacha (ESP17_3_TIQUI)

"He$_i$ tries to bring her$_j$ some joy because she$_j$ suffers from a disease and she$_j$ does not remember anything of her$_j$ past, not even of her$_j$ own family. **He$_i$** starts to tell her$_j$ a story about a girl"

-4plusclauses

105) Ella$_i$ teine una hermana$_j$. El nombre de la hermana$_j$ menor es Ashlee Simpson. Ashlee$_j$ es una cantante tambien y Ø$_j$ es bonita. **Jessica$_i$** rompó a Nick Lachey$_k$ en el ano pasado (ENG24_19_6_2_SH)

"She$_i$ has a sister$_j$. The name of the little sister$_j$ is Ashlee Simpson. Ashlee$_j$ is also a singer and she$_j$ is pretty. **Jessica$_i$** broke up with Nick Lachey$_k$ last year"

5.5.2.4 Antecedent syntactic function

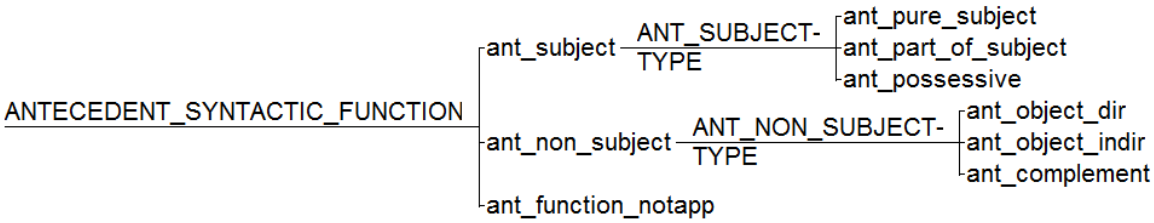


Figure 13. Antecedent syntactic function

The syntactic function of the antecedent has been extensively investigated in the PAS studies on AR (Alonso-Ovalle, Fernández-Solera, Frazier, & Clifton, 2002; Bel & García-Alcaraz, 2015; Belletti, Bennati, & Sorace, 2007; Carminati, 2002; Jegerski, VanPatten, & Keating, 2011; Keating, VanPatten, & Jegerski, 2011; Kras, 2008; Papadopoulou, Peristeri, Plemenou, Marinis, & Tsimpli, 2015; Tsimpli & Sorace, 2006; Tsimpli, Sorace,

Heycock, & Filiaci, 2004). Regarding Spanish language, evidence suggests that when the antecedent is in subject position, it is more likely to be recovered with less explicit referential forms, whereas no such preference is reported for object antecedents. Notice, however, that the above evidence comes exclusively from experimental data, where only a very specific AR pattern has been tested, namely: two 3rd person singular same gender referents, one in subject and the other in object position of a main clause, followed by an anaphor in the exact next clause. The PAS account is thus restricted to represent only "one narrow instance of the numerous and complex set of discourse structure variables" (Jegerski et al., 2011:503). Consider, for example, the case represented in example (98) where the antecedent is only part of a plural subject form. Alternatively, it may be only part of a possessive noun phrase in subject position, as in example (95). In this study, in order to account for the wider range of anaphoric patterns present in real discourse, all anaphors were analytically tagged for the exact syntactic function of the antecedent. All possible antecedent forms and syntactic functions are being accounted for, as it can be seen in the following examples extracted from the corpus:

-ant_subject/ant_pure_subject

106) Cuando él$_i$ era joven, **él$_i$** quería ser futbolista (ENG23_23_3_2_JP)
"When he$_i$ was young, **he$_i$** wanted to be a football player"

-ant_subject/ant_part_of_subject

107) Ø$_{ij}$ estan bailando durante toda la noche y Ø$_{ij}$ se enamoran. Pero **ella$_i$** tiene que irse (GR37_24_3_3_NAT)
"They$_{ij}$ are dancing during the whole evening and they$_{ij}$ fall in love. But **she$_i$** has to go"

-ant_subject/ant_possessive

108) La perspectiva de Rafael$_i$ cambia cuando **Ø$_i$** sufre un infarto (ESP21_3_TW)
"Rafael$_i$´s perspective changes when **he$_i$** has a heart attack"

-ant_non_subject/ant_object_dir

109) Ellis$_i$ decide invitar a Adaline$_j$ al universario de sus$_i$ padres. Aquí **Ø$_j$** se encuentra con su$_i$ padre William Jones (GR38_21_5_3_IFI)
"Ellis$_i$ decides to invite Adaline$_j$ to his$_i$ parents´ anniversary. There **(she)$_j$** meets his$_i$ father William Jones"

-ant_non_subject/ant_object_indir

110) Su$_i$ hermana$_j$ le$_i$ ayuda a encontrar un trabajo a una iglesia y allí **Ø$_i$** conoce a algunas mujeres fuertes (ENG41_19_5_3_AEM)

"Her$_i$ sister$_j$ helps her$_i$ to find a job in a church and there **(she)$_i$** meets some strong women"

-ant_non_subject/ant_complement

111) Ø$_i$ está casado con alemana-mexicana Monique Obermuller$_j$ desde 2004. **Obermuller$_j$** es una actriz (GR34_42_1_2_YOR)

"(He)$_i$ is married to the German-Mexican Monique Obermuller$_j$ since 2004. **Obermuller$_j$** is an actress"

-ant_function_notapp

112) Pero él$_i$ se mantuvo a su$_j$ lado hasta que al final **Ø$_{ij}$** murieron juntos (ESP17_3_TIQUI)

"But he$_i$ stayed by her$_j$ side until at the end **they$_{ij}$** died together"

5.5.2.5 PAS in discourse

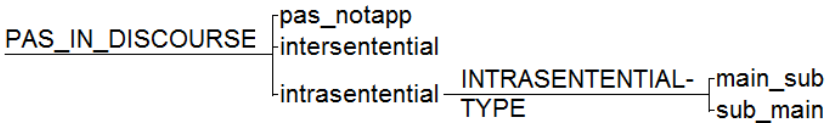


Figure 14. PAS in discourse

In order to thoroughly test the well-studied PAS structure in discourse (see previous section), all referential subject expressions of this study were further tagged for pertaining to such a syntactic structure or not. This will allow full comparability with the results of the previous literature on this area. In line with the experimental studies, two anaphoric patterns that have been widely tested in the literature (Carminati, 2002; Filiaci, Sorace, & Carreiras, 2014; Keating, VanPatten, & Jegerski, 2011; Sorace & Filiaci, 2006) were distinguished: intersentential and intrasentential. In the first case, the anaphor is located in a main clause, whereas in the second case, it is located in a subordinate clause. In the second case, the order of main and subordinate clause was further accounted for. It is important to keep in mind that the PAS structure considers anaphoric patterns where only two-same gender referents are implicated, one in subject and the other in object position. Although in the experimental studies the two referents are typically introduced as full noun phrases, a more flexible operationalization was adopted in this study: the two potential antecedents could also be either null or overt pronouns. All non-relevant cases were tagged as not applicable. Examples of each category of the PAS structure are given below:

-intersentential

113) Neruda$_i$ y su$_i$ poesía fue muy importante porque eventualmente él$_i$ ayudó Mario$_j$ mucho con su amor. Primero, **Mario$_j$** aprendió sobre la poesía (ENG35_15_7_3_LMR)

"Neruda$_i$ and his poetry was very important because eventually he$_i$ helped Mario$_j$ a lot with his love. First, **Mario$_j$** learned about poetry"

-intrasentential/main_sub

114) Wheeler$_i$ trajo Ronnie$_j$ a una fiesta con muchas chicas bellas porque **Ronnie$_j$** se amo chichas (ENG30_20_6_3_NJP)

"Wheeler$_i$ brought Ronnie$_j$ to a party with a lot of girls because **Ronnie$_j$** loved girls"

-intrasentential/sub_main

115) Cuando Freddie$_i$ se enamora de la hija$_j$ de su jefe que esta comprometida con otro, **Ø$_i$** descubre que cometio un error (GR41_39_7_3_FOT)

"When Freddie$_i$ falls in love with the daughter$_j$ of his boss who is engaged to another guy, **he$_i$** discovers that he made a mistake"

5.5.2.6 Protagonist

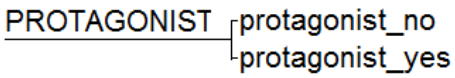


Figure 15. Protagonist

The Protagonist feature accounts for the fact that some referents in a text are a priori more prominent than others. According to Huang (2000:154): "there has been some general consensus in the literature that the protagonist enjoys a special thematic status in a narrative or conversation, thus frequently receiving a minimal anaphoric encoding after initial introduction". This is further in line with the notion of Discourse Topic in Van Dijk (1977) defined "in terms of repeated reference to a given discourse referent" (p.56). The same idea appears in numerous studies under very similar definitions (Chafe, 1994; Clancy, 1980; Givón, 1990; Grimes, 1978; Ryan, 2015). In the same line, the label Protagonisthood has been used to express this feature in Kibrik (2000), According to the author, "it specifies whether the referent is the main character of the discourse" (p.78). In order to account for the well-attested effect of this status in discourse, all referents were tagged under the label Protagonist. Due to the lack of objective criteria for the identification of the protagonist for each text, two explicit instructions were followed. In the expository essays ("Write about a famous person") only the main character was considered, as in the example below:

-protagonist_yes

116) Barack Obama$_i$ está el persona más fomoso en el mundo hoy. **Él$_i$** está el presidente de las Estadoas Unidos (ENG22_18_5_2_KAC)

"Barack Obama$_i$ is the most famous person in the world today. **He$_i$** is the president of the United States"

In the narrative essays ("Summarize a film that you have watched recently") the protagonist was usually defined by the author in some point of the narration, as in the example below:

-protagonist_yes

117) **El protagonista** se enamora con una mujer (GR37_38_2_3_DIM)
"**The protagonist** falls in love with a woman"

All other referents were tagged as not being protagonists.

5.5.2.7 New paragraph

NEW_PAR
- new_par_no
- new_par_yes

Figure 16. New paragraph

Paragraph boundary has been traditionally recognized as a relevant factor for anaphora in written discourse. Hinds (1977) was the first to show that "paragraph structure influences the appearance or nonappearance of pronouns" (p.95). In line with this, Hofmann (1989) claims that "a pronoun or other anaphoric element cannot be used if its nearest antecedent is embedded in a preceding paragraph" (p.241). Similar observations are made by Tomlin (1987:29) and Fox (1987:113). Huang (2000) further notes that "mentions (initial or non-initial) at the beginning or peak of a new discourse structural unit tend to be encoded by a full NP" (p.172). Lozano (2016) further confirms this observation in his corpus study: "An NP can thus corefer with the subject antecedent (topic) in the last sentence of the preceding paragraph" (p.266). In order to examine to what extent the New Paragraph factor is a barrier to anaphora, all subject forms were tagged for starting a new paragraph or not, as in the example below:

-new_par_yes

118) Como muchos saben, Rafa Nadal$_i$ es uno de los mejores tenistas del momento y quizás en un futuro, de la historia.
Rafa Nadal$_i$ es un tenista de unos 24 años nacido en Manacor (ESP26_2_AVS)
"As many people know, Rafa Nadal$_i$ is one of the best tennis players right now and maybe in the future, of the whole history.
Rafa Nadal$_i$ is a 24 years old tennis player born in Manacor"

5.5.2.8 Active referents

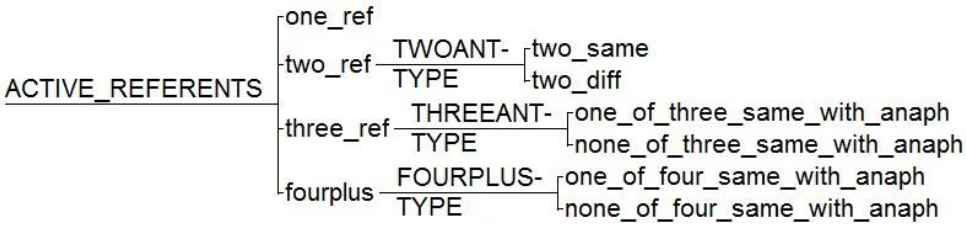


Figure 17. Active referents

As we saw in 2.4.2.1, one of the three factors used by Givón (1983) to measure topicality in discourse is Potential Interference (the other two being Distance and Persistence). This measurement aims to assess "the disruptive effect which other referents within the immediately preceding register may have" (p.14). The same factor was reformulated by Sun & Givón (1985) under the label Potential Referential Interference (PRI) which is defined as "the number of other referents in the directly preceding discourse environment – most commonly 3 clauses – that are semantically compatible with the predicate of the referent under consideration" (p.331). The idea was further refined by Fox (1987) who distinguished between three contexts of pronominalization: no interference, different-gender referents, and same-gender referents. Her evidence demonstrates the "widespread use of full NP in the same-gender environment for the written material" (p.147). The PRI factor was later applied to corpus data (Flores-Ferrán, 2002, 2004; Lozano, 2016; Travis & Cacoullos, 2012) and was also found to be significant in psycholinguistic experiments (Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000; Arnold & Griffin, 2007). In order to fully operationalize the PRI factor, following Fox (1987) and Lozano (2016), same-gender and different-gender contexts will be distinguished. Potential interference is calculated following the definition of Sun & Givón (1985). All semantically compatible referents between the sentence where the antecedent appears and the anaphor under study are considered. Examples for each category are provided below:

-one_ref

119) Adaline$_i$ sale de la casa porque **Ø$_i$** se siente que su secreto corre peligro (GR38_21_5_3_IFI)

"Adaline$_i$ gets out of the house because **(she)$_i$** feels that her secret is in danger"

-two_ref/two_same

120) Antón Chigurh$_i$ tiene muchos careos con el cazador$_j$ con la moneda, pero, Ø$_i$ no pudo matar el cazador$_j$. **Antón$_i$** usó muchas armas (ENG25_66_4_3_HR)

"Anton Chigurh$_i$ has many confrontations with the hunter$_j$ with the coin, but, he$_i$ could not kill the hunter$_j$. **Anton$_i$** used a lot of weapons"

-two_ref/two_diff

121) El hombre$_i$ propone a su novia$_j$ participar en una pelicula porno para ganar dinero. La mujer$_j$ esta de acuerdo y Ø$_{ij}$ lo hacen, sin que este acontecimiento afecte a su$_{ij}$ relacion. **El hombre$_i$** promete comprarle$_j$ una casa bonita (GR43_46_2_3_TIM)

"The man$_i$ proposes to his girlfriend$_j$ to play in a porn film in order to get some money. The woman$_j$ agrees and they$_{ij}$ do it, without any consequences for their$_{ij}$ relationship. **The man$_i$** promises to buy her$_j$ a beautiful house"

-three_ref/one_of_three_same_with_anaph

122) La primera opción básicamente se trata de un triángulo amoroso entre Carlos$_i$, Julia$_j$ y Pedro$_k$. **Carlos$_i$** era casado con Julia$_j$ (ENG41_28_15_3_KDH)

"The first option was basically a love triangle between Carlos$_i$, Julia$_j$ and Pedro$_k$. **Carlos$_i$** was married to Julia$_j$"

-three_ref/none_of_three_same_with_anaph

123) El otro dia el principe$_i$ la$_j$ esta buscando, pero su$_j$ madrastra$_k$ no la$_j$ permite que Ø$_j$ se vaya de la casa. Despues de muchas problemas, **el principe$_i$** decide que todas las mujeres del reino pueden probar el zapatillo (GR37_24_3_3_NAT)

"The other day the prince$_i$ is looking for her$_j$, but her$_j$ stepmother$_k$ does not let her$_j$ leave the house. After many problems, **the prince$_i$** decides that every woman in the kingdom can try the shoe"

-fourplus_ref/one_of_four_same_with_anaph

124) Ronnie$_i$ oír por casualidad la conversación y Ø$_i$ decide a perdonar Wheeler$_j$ pero Danny$_k$ necesito hacer más pero recibir la confianza de Augie$_l$. **Danny$_k$** hablo con el rey del mundo imaginario (ENG30_20_6_3_NJP)

"Ronnie$_i$ accidentally hears the conversation and (he)$_i$ decides to forgive Wheeler$_j$ but Danny$_k$ needs to do more in order to receive the confidence of Augie$_l$. **Danny$_k$** talked with the king of the imaginary world"

-fourplus_ref/none_of_four_same_with_anaph

125) Rufus$_i$ y su hermana, Martha$_j$, se quedaron con su madre, Kate$_k$. Luego, Kate$_k$ se fue a Montreal, en Canadá, para estar con su hermana, Anna$_l$, con quién Ø$_k$ hacia su$_k$ música. Por eso, **Rufus$_i$** pasaba sus$_i$ años principales en Montreal (ENG37_17_9_2_CJR)

"Rufus$_i$ and his sister, Martha$_j$, stayed with their mother, Kate$_k$. Later, Kate$_k$ went to Montreal, Canada, to be with her$_k$ sister, Anna$_l$, with whom (she)$_k$ made her$_k$ music. That´s why **Rufus$_i$** passed his time mostly in Montreal"

5.5.2.9 Shared knowledge constraints

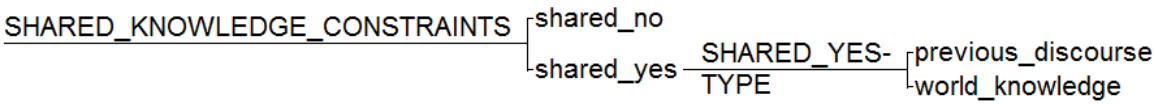


Figure 18. Shared knowledge constraints

The role of context has been consistently highlighted in previous studies on discourse anaphora. In words of Prince (1981) the relevant question concerning discourse production is "what kinds of assumptions about the hearer/reader have a bearing on the form of the text being produced" (p.233). Givón (1983:16) accounts for this assumptions by distinguishing between three types of shared knowledge:

i. Generically shared knowledge coded in the culturally shared lexicon (world knowledge)
ii. Specifically shared knowledge of the particular discourse (discourse knowledge)
iii. Specifically shared knowledge of the particular speaker and hearer (personal knowledge)

Blackwell (1998:614) cites Clark & Marshall (1981) who define mutual knowledge as "*A* knows that *A* and *B* mutually know *p*"(p.18). She concludes that "given this definition, different types of shared background knowledge (cultural, social, stereotypical, and discursive) might be viewed as subsets of mutual knowledge" (p.614). This is further in line with the neo-Gricean approaches of Levinson (1991, 1995) and Huang (2000a). The same idea also appears in Emmott (1997, 2006) under the labels *schema knowledge* and *text world knowledge*, whereas Eslami Rasekh (1997) uses the term *script* for "things mutually known to participants of the discourse communication" (p.35). Finally, Arnold, Kaiser, Kahn, & Kim (2013) refer to *common ground* which includes "a social or cultural background, a linguistic or environmental domain, and expectations about the course of the conversation" (p.411). In short, in written discourse where the writer does not have a personal relationship with the eventual reader, only two of the three Shared Knowledge Constraints may be considered. The first relates to the previous discourse (schema, script, text world, linguistic domain, etc.) and the second accounts for the world knowledge (cultural, social, stereotypical, etc.). It should be noted here that this is the only factor of the tagset that may allow for some subjective interpretation. For that reason, a rather conservative analysis was adopted regarding this constraint, insofar as only strongly

marked cases were tagged as being or not constrained by either previous discourse or world knowledge. Consider the example below:

-shared_yes/previous_discourse

126) La perspectiva de Rafael$_i$ cambia cuando Ø$_i$ sufre un infarto a causa de su$_i$ agotamiento nervioso y por un tiempo Ø$_i$ no puede trabajar. Ese hecho coincide con el encuentro con un amigo$_j$ de la secundaria, al que Ø$_i$ no había visto por veinte años, que es actor y Ø$_j$ tiene mayor sensibilidad. **Ø$_j$** lo$_i$ convence de que **Ø$_i$** acepte la propuesta de su$_i$ padre y cuando **Ø$_i$** se mejora, **Ø$_j$** empieza a ayudarlo$_i$ con los preparativos. (ESP21_3_TW)

"Rafael$_i$´s perspective changes when he$_i$ has a heart attack because of his$_i$ nervous breakdown and for a while he$_i$ cannot work. This fact coincides with a reunion with a friend$_j$ from high school, whom (he)$_i$ had not seen for twenty years, who is an actor and (he)$_j$ has more sensibility. **(He)$_j$** convinces him$_i$ to accept the proposal of his$_i$ father and when **(he)$_i$** gets better, **(he)$_j$** starts helping him$_i$ with the preparations"

In example (126), the null subjects under consideration would be insolvably ambiguous if it was not already known from previous discourse that Rafael's father made a proposal to his son. This information makes it logical to assume that the friend is the one who "convinces" Rafael to "accept the proposal" and not the other way round. We also know that Rafael is the one who had a heart attack, so we can assume that he is the one who "gets better" and his friend is the one who "starts helping him". Consider now how world knowledge may also decisively constraint referential choices in the following example:

-shared_yes/world_knowledge

127) John Lennon$_i$ nació en Inglaterra el día 09 de Octubre de 1940, bajo un ataque aéreo de la armada alemana. Su$_i$ padre$_j$ los$_{ik}$ abandonó cuando **él$_i$** era muy niño (ESP44_2_ASJ)

"John Lennon$_i$ was born in England on the 9th of October of 1940, under an aerial attack of the German army. His$_i$ father$_j$ abandoned them$_{ik}$ when **he$_i$** was still a child"

In (127), the pronominal form in question may refer both to John Lennon and to his father. In order to understand how this might be, imagine an alternative, but very similar sentence: "His$_i$ father$_j$ abandoned them$_{ik}$ when **he$_j$** fell in love with another woman". Note that only world knowledge may resolve the anaphoric relation in both the real and the imaginary example. In the example (127), "he" refers to John Lennon, because we know that a father may abandon his son when the boy is "still a child". Assuming that John Lennon abandoned his father when the latter was still a child would be irrational. Similar assumptions can be made regarding the imaginary alternative example, where the same overt pronominal corefers here with the father and not the son for the simple reason that it makes more sense according to world knowledge: a father may abandon his family

when he falls in love with another woman (but not when the son falls in love with another woman).

5.5.3 Pragmaticality

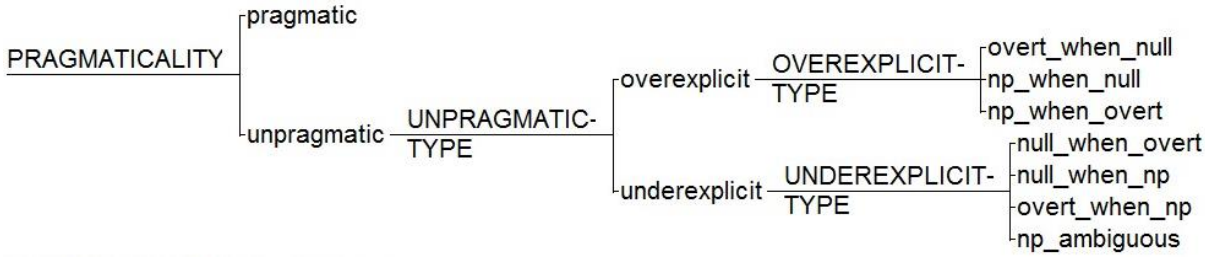


Figure 19. Pragmaticality

In order to account for potential differences between groups regarding the felicitous choice of anaphoric expression, all referential forms were further tagged for being pragmatic or unpragmatic. Recall here that, essentially, all referential choices are grammatical. However, not all of them are pragmatically appropriate in all contexts. Following previous LCR studies (Lozano, 2009b, 2016; Montrul, 2004a; Montrul & Rodríguez Louro, 2006) two types of illicit use of referential expressions were identified: overexplicitness (redundancy) and underexplicitness (ambiguity). Following Lozano (2016:15) overexplicit and underexplicit referential forms were further classified to another three and four subtypes respectively. Due to the lack of a straightforward operationalization for Pragmaticality in the literature (see section 2.4.3) and in order to ensure comparability among groups, several discourse patterns were explicitly defined for each subtype of unpragmatic use. More specifically:

i. When there is no switch in reference, an overt pronoun or a noun phrase are considered overexplicit (unless when starting a new paragraph), as in the examples below:

   -unpragmatic/overexplicit/overt_when_null

   128) José Antonio Domínguez Banderas$_i$ es un actor, cantante y productor de cine español. **El$_i$** nació el 10 de agosto de 1960 en una ciudad pequeña en Málaga. **Él$_i$** tiene 54 años (GR21_22_1_2_JUA)

   "Jose Antonio Dominguez Banderas$_i$ is an actor, singer and producer of Spanish cinema. **He$_i$** was born at 10th of August of 1960 in a small city in Malaga. **He$_i$** is 54 years old"

-unpragmatic/overexplicit/np_when_null

129) Mario$_i$ creyó que él$_i$ fue el problema porque Ø$_i$ nunca hizo algo importante en su$_i$ vida. Por eso **Mario$_i$** escribí una poema (ENG35_15_7_3_LMR)

"Mario$_i$ thought that he$_i$ was the problem beacause (he)$_i$ never did anything important in his$_i$ life. That´s why **Mario$_i$** wrote a poem"

ii. When there is switch in reference, but only one referent is active, an overt pronoun or a noun phrase are considered overexplicit (unless when starting a new paragraph), as in the example below:

-unpragmatic/overexplicit/overt_when_null

130) La vida de Barack Obama$_i$ es muy interesante. **Él$_i$** nació en Hawaii (ENG22_18_5_2_KAC)

"The life of Barack Obama$_i$ is very interesting. **He$_i$** was born in Hawaii"

-unpragmatic/overexplicit/np_when_null

131) Ella$_i$ tiene que irse hasta las doce de medianoche, porque despues los hechizos se van. Por eso, **Cenicienta$_i$** abandona la fiesta (GR37_24_3_3_NAT)

"She$_i$ has to go before midnight because after that the spell is gone. That´s why **Cinderella$_i$** leaves the party"

iii. A noun phrase in contexts with two or more different-gender active referents (where an overt pronoun would be sufficiently informative) is considered overexplicit (unless it starts a new paragraph), as in the example below:

-unpragmatic/overexplicit/np_when_overt

132) Al poco tiempo su$_i$ novia$_j$ se cansa de la falta de atención y Ø$_j$ le$_i$ dice que Ø$_j$ quiere cortar la relación. En ese momento **Rafael$_i$** se da cuenta de que Ø$_i$ la$_j$ quiere (ESP21_3_TW)

"Very soon his$_i$ girlfriend$_j$ is sick of the lack of attention and (she)$_j$ tells him$_i$ that she$_j$ wants to break the relationship. In that moment **Rafael$_i$** realizes that (he)$_i$ wants her$_j$"

iv. A null pronoun in switch-reference contexts with more than one different-gender active referents is considered underexplicit, as in the example below:

-unpragmatic/underexplicit/null_when_overt

133) Todo estaba excelente en su$_i$ vida cuando Ø$_i$ vio en un periódico la foto de Noah$_j$ con la casa hermosa que **Ø$_j$** le$_i$ prometió que le$_i$ iba a construir (ESP17_3_TIQUI)

"Everything was excellent in her$_i$ life when she$_i$ saw in the newspaper the photo of Noah$_j$ with the beautiful house that **(he)$_j$** promised that (he)$_j$ would build her$_i$"

v. A null pronoun in switch-reference contexts with more than one same-gender active referents is considered underexplicit, as in the example below:

-unpragmatic/underexplicit/null_when_np

134) Mientras Ø$_i$ arregla el tejado de una casa, Ø$_i$ se entera casualmente de que el dueño$_j$ de la casa iba a participar en un asunto que le$_j$ iba a proporcionar una gran cantidad de dinero. El dueño$_j$ de la casa muere, su$_j$ compañera$_k$ no puede pagar los servicios de saneamiento del tejado que **Ø$_i$** había prestado (ESP32_3_MDD)

"While (he)$_i$ fixes the roof of a house, (he)$_i$ accidentaly discovers that the owner$_j$ of the house was going to participate in a matter that would provide him$_j$ a big sum of money. The owner$_j$ of the house dies, his$_j$ girlfriend$_k$ cannot pay the services of the roof that **(he)$_i$** had offered"

vi. An overt pronoun in any context with more than one same-gender active referents is considered underexplicit, as in the example below:

-unpragmatic/underexplicit/overt_when_np

135) Empieza cuando ella$_i$ se va a la casa de su$_i$ hermana$_j$ después de una noche del abuso, suponemos. **Ella$_j$** nunca dice directamente que está pasando en su$_j$ vida (ENG41_19_5_3_AEM)

"It starts when she$_i$ goes at her$_i$ sister$_j$´s house after a night of abuse, we suppose. **She$_j$** never says directly what is going on in her$_j$ life"

vii. Finally, some very infrequent cases of ambiguous noun phrases were encountered and tagged, as in the example below:

-unpragmatic/underexplicit/np_ambiguous

136) Esa mujer$_i$ era diabolica y Ø$_i$ tenia dos hijas$_j$ que tambien eran muy malas frente de Cinderella$_k$. Cuando **su$_{k/j}$ padre$_1$** murio cuando Ø$_1$ viajaba, la matriz$_i$ comenzó manipular Cindirella$_k$ (GR35_19_4_3_SS)

"This woman$_i$ was diabolic and (she)$_i$ had two daughters$_j$ that were also very mean to Cindirella$_k$. When **her$_k$/their$_j$ father$_1$** died when (he)$_1$ was travelling, the doting mother$_i$ started to manipulate Cindirella$_k$"

5.6 The final dataset

The full annotation scheme was manually implemented to the 92 texts of the corpus sample (see Table 2). Each 3rd person subject of a tensed clause was tagged according to the features of the pair anaphor/antecedent and their pragmatic felicity/infelicity, as described in the previous section. It should be noted here that, in accordance with the previous literature on anaphora in Spanish discourse, categorical contexts were excluded from the analysis. The so-called 'envelope of variation' (Blackwell & Quesada, 2012; Cameron, 1994, 1995; Cameron & Flores-Ferrán, 2004; Geeslin & Gudmestad, 2016; Gudmestad & Geeslin, 2010; Otheguy & Zentella, 2012; Otheguy, Zentella, & Livert, 2007) refers to the linguistic contexts in which the variable under study may vary. In this

study, the contexts that were found to allow only one of the possible subject forms (and were, thus, excluded from the analysis) are:

i. Expletive clauses: existential and weather verbs

137) **Ø** Es interesante decir que la película no en colores (ENG37_22_2_3_DH)

"**(It)** is interesting to say that the film in not in color"

ii. Subject relative clauses

138) Una pareja joven, de unos veinte años, que **Ø** vive en Madrid (GR43_46_2_3_TIM)

"A young couple, around the age of twenty, which lives in Madrid"

iii. Impersonal and passive clauses

139) Cuando **Ø** se discute sobre las causas de su ejecución (GR43_22_1_2_ATH)

"When **(it)** is being discussed about the reasons of his execution"

All the above constructions do not allow the alternation of anaphoric subject forms and null forms were exclusively employed by all participants in these contexts. Therefore, the subjects of these verbs were excluded from the annotation process. The total number of texts, words and tagged items per group is summarized in Table 10:

Group	#texts	#words	#tagged items
Natives	20	6.875	501
English1	12	3.869	409
English2	12	4.208	359
English3	12	7.318	507
Greek1	12	3.576	272
Greek2	12	3.511	308
Greek3	12	4.551	295
TOTAL	**92**	**33.908**	**2.651**

Table 10. Summary of annotated data for each group

As can be seen, a total of 2.651 subject forms were tagged. Considering that each item was manually tagged for 15 features (see the full annotation scheme in Figure 53 in the Appendix), the total number of manually assigned tags exceeds 39.750[118].

Finally, each text was further annotated according to the group where it belongs, as can be seen in Figure 20:

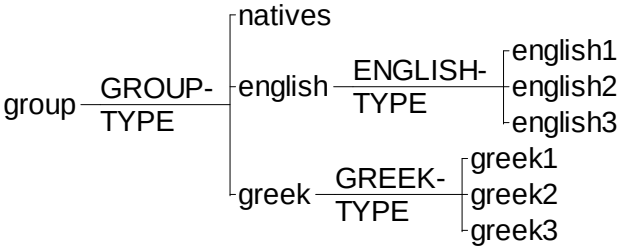


Figure 20. Annotation of texts according to group

In sum, this section has focused on methodological issues regarding the steps that were followed in order to perform the analysis of the data. The aim was to provide a detailed description of the corpus, the participants, the software and the tagsets that were used in this study. Note here that the combination of the two annotation schemes (Figure 20 and Figure 53) gives access to all sorts of inter- and intra-group comparisons for any of the properties of the annotated subject expressions. Several kinds of intricate search queries that may serve for contrastive purposes were performed and will be presented in the Results chapter that follows.

[118] For some categories, more than one tag had to be manually assigned, e.g. in the Clause Type category, a coordinate cumulative clause was first tagged as 'coordinate' (for clause type) and subsequently as 'cumulative' (for type of coordinate clause).

CHAPTER 6

6 RESULTS AND DISCUSSION

In this chapter, the results of this study will be presented, analysed and discussed. The first section of the results deals with the overall distribution of subject forms (null, overt pronoun, NP, demonstrative, other) for the entire dataset. Although of purely descriptive nature, the overall distributions of the referential choices for each group may provide some preliminary insights regarding general trends which will be examined in depth during the statistical analysis. Subsequently the native control group is examined separately, with the purpose of giving a full account regarding the factors that affect referential choices in Spanish L1. Finally, the main section of the results focuses on the pragmaticality of the referential choices, starting with an overview of the unpragmatic forms for all groups together. Thenceforward, a contrastive interlanguage analysis (CIA) is analytically performed, starting with the intermediate proficiency groups (English1 and Greek1), followed by the advanced (English2 and Greek2) and upper-advanced (English3 and Greek3) groups. All learner groups are first contrasted with the native control group and then separately analysed, with the intention of providing both a contrastive analysis and a complete account of each interlanguage in its own right regarding the pragmatic felicity of the produced anaphoric subjects. After that, an overall comparison of all groups together is performed in order to fully account for developmental issues.

During the analysis, inferential statistics (two-tailed chi-square tests with Yate's correction) were performed with the significance level maintained at 5%. Fisher's exact tests were used when the observed raw frequencies were equal or less than five. All the original search queries, performed in the UAM CorpusTool, as well as the resulting raw frequencies and percentages can be visualized in the Appendix. Due to the fact that UAM CorpusTool reports only chi-square values (but not the exact p value), all pairwise statistical comparisons were performed with the GraphPad software[119] and the Chi-square Calculator[120]. In the case of larger than 2x2 tables, for all post-hoc pairwise comparisons, the Bonferroni adjustment was used in order to control maximum type 1 error rate and

[119] https://graphpad.com/

[120] http://turner.faculty.swau.edu/mathematics/math241/materials/contablecalc/

the significance level was adjusted according to the total number of pairwise comparisons (Beasley & Schumacker, 1995; Garcia-Pérez & Nuñez-Anton, 2003; McDonald, 2014). Results of each statistical comparison are briefly discussed in situ and further reviewed in the summary of each section with respect to the proposed Hypotheses in Chapter 4.

6.1 Overall distribution of subject expressions

As we already saw in section 5.6, there is a total number of 2.651 3rd person subject expressions in the dataset. A very important distinction between anaphoric (previously mentioned in the text) and non-anaphoric subjects (first-time mentioned in the text) needs to be made here before proceeding with the analysis. Crucially, all the referents that are mentioned for the first time in a text do not corefer with any previously mentioned referent and thus cannot be tagged for antecedent features. Note, additionally, that a first-mentioned referent is typically introduced with a noun phrase since it represents discursively new information[121]. Therefore, the presence of the first-mentioned referents in the analysis, as being directly comparable to anaphoric subjects (defined in terms of coreference with a textual antecedent), may severely skew the results. Consequently, for comparability reasons, the first-mentioned subjects were extracted from the dataset in order to be analysed separately in the future. Similarly, the very few cases of insolvably ambiguous anaphoric forms (a total number of 8 anaphoric subjects accounting for the 0.3% of the data) were also removed since no anaphoric relation could be established and annotated in this case. After removing the non-anaphoric first-mentioned and the insolvably ambiguous items, 2.060 3rd person anaphoric subjects remained. The remaining anaphors are distributed by group and form as shown in Table 11 (the original UAM CorpusTool raw frequencies can be seen in Figure 54 in the Appendix):

[121] Otherwise it would be insolvably ambiguous (at least in the discourse genre under study).

group / form	Natives		English1		English2		English3		Greek1		Greek2		Greek3	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
null	219	**59.51**	82	**25.39**	117	**41.64**	209	**49.18**	141	**69.80**	172	**69.35**	133	**62.44**
overt	26	**7.07**	105	**32.51**	55	**19.57**	65	**15.29**	16	**7.92**	11	**4.44**	9	**4.23**
NP	121	**32.88**	134	**41.49**	107	**38.08**	149	**35.06**	45	**22.28**	63	**25.40**	71	**33.33**
dem.	2	**0.54**	0	**0.00**	1	**0.36**	1	**0.24**	0	**0.00**	0	**0.00**	0	**0.00**
other	0	**0.00**	2	**0.62**	1	**0.36**	1	**0.24**	0	**0.00**	2	**0.81**	0	**0.00**
Total	368	**100**	323	**100**	281	**100**	425	**100**	202	**100**	248	**100**	213	**100**

Table 11. Overall distribution of anaphoric subjects per group

Overall distributions have been widely employed in the literature for contrasting purposes. However tempting this might be, I will argue together with Geeslin & Gudmestad (2016:65) that overall distributions should be strictly limited to general observations since they do not meet the assumptions for statistical comparison (the data are not normally distributed). As Ryan (2015:849) points out, "it could be misleading to measure overexplicitness through straightforward comparisons of the relative number of pronoun and zero tokens in NS and L2". In other words, an elevated proportion of overt subjects does not entail overexplicitness until the pragmatic felicity of the anaphors in specific discourse patterns has been determined (see also section 2.4.3). That being said, and despite the merely descriptive nature of this kind of data, some general observations concerning potential differences and similarities between groups can be made and will be pointed out.

The first observation regarding the overall distribution of anaphoric subjects in Figure 21 concerns the very scarce quantity of demonstrative and other pronouns in the dataset. The two categories together barely represent the 0.48% (10 cases) of the total number of annotated subjects for the entire dataset. This might be a discourse genre-specific phenomenon, since "demonstrative pronouns are most commonly used to refer to abstract entities in Spanish" (Zulaica-Hernández, 2016:20). Written narrative discourse with a very high interaction of animate referents seems to disfavour the use of demonstratives and other secondary anaphoric pronouns. Given that no conclusions may be drawn from such a low number of cases, the above two categories were removed from the dataset in order to simplify the analysis and allow comparability with previous studies that have mainly examined the three most common types (null, overt pronoun, noun phrase). Thus,

for the rest of the analysis we will focus exclusively on these three prototypical anaphoric subject expressions which account for the 99.57% of the data.

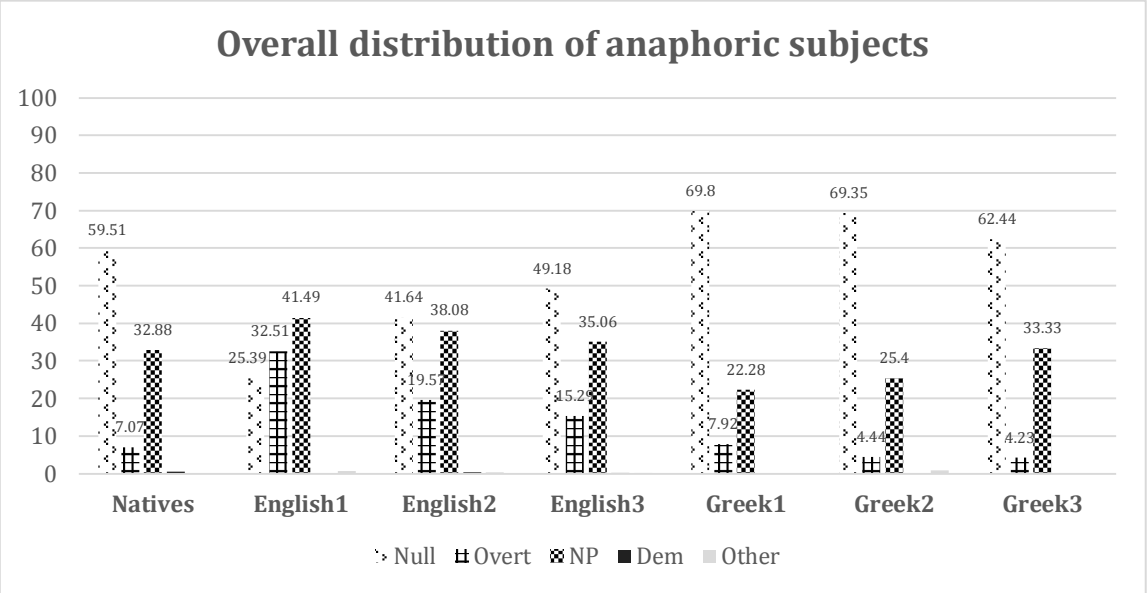


Figure 21. Overall distribution of anaphoric subjects per group

The second observation concerns the numerous NP anaphoric subjects which are employed indistinctively by all groups. Note here that overt pronouns have been traditionally considered as one of the two anaphoric options in Spanish (the other one being null subjects). The alternation between null and overt pronouns has been widely examined from different theoretical perspectives and with different methodologies. On the other hand, noun phrases have been systematically ignored in the vast majority of previous studies (some exceptions are: Blackwell & Quesada, 2012; Dumont, 2006; Gudmestad, House, & Geeslin, 2013; Lozano, 2009b, 2016). Following these studies, the present thesis argues for the need to include NPs in the analysis of 3rd person anaphoric expressions. It should be noted that the results of all studies that have considered NPs in the analysis coincide in a crucial finding: for 3rd person anaphoric subjects, in terms of frequency of use, the noun phrases (and not the overt pronominals) constitute the major alternative to null subjects in Spanish (see also 3.1). Therefore, the exclusion of NPs from any study on 3rd person anaphora could severely bias the results of the analysis.

Finally, regarding the overall distribution of the three main anaphoric subject forms per group presented in Figure 21, we observe largely the same pattern for all groups (with the exception of the English1 participants): overall, the majority of the subjects are unexpressed (null), and their frequency is followed by an important proportion of noun phrases. Overt pronouns are the type of anaphoric expression which is less employed by

all groups. A quite different pattern is observed for the English1 group: more noun phrases, followed by an important proportion of overt pronouns and relatively few null subjects. Recall here that these are mere observations since, as already argued, statistical tests may not be validly performed for overall distributions.

In sum, the overall distribution of the data provides some preliminary insights regarding possible variability between groups in the usage of anaphoric expressions. More specifically, the distributional pattern of the English-speaking learners seems to differ from that of their Greek counterparts and the control group participants. Overall, they seem to employ to a greater extent overt anaphoric expressions (pronouns and noun phrases). Greek-speaking learners, on the other hand, also seem to differ from the control group, although in the opposite direction: they seem to favour null subjects, even more than the Spanish natives. However, Greek2 and Greek3 groups seem to present a relatively similar distributional pattern to the native control group.

6.2 Anaphora factors in Spanish L1

Although the main focus of this study is on the interlanguage of learners of L2 Spanish, the annotated data may also provide some insights regarding the anaphoric production of the Spanish native speakers. Therefore, we shall begin the analysis by focusing on the referential choices of the Spanish control group. Several discourse factors that have been claimed to account for the production of anaphoric subjects in Spanish will be briefly examined here. Recall that, as it has been argued, there are three major referential forms in Spanish language: null subjects, overt pronouns and lexical noun phrases (see section 3.1). On the one hand, null anaphors are the less specified forms of reference and constitute the most frequent referential choice in our dataset. On the other hand, overt subjects (pronouns and noun phrases) are more specified and overall less frequent referential forms. In order to simplify the analysis of the Spanish L1 data, the very few overt pronouns of the dataset were temporarily merged with the noun phrases, as can be seen in Figure 22:

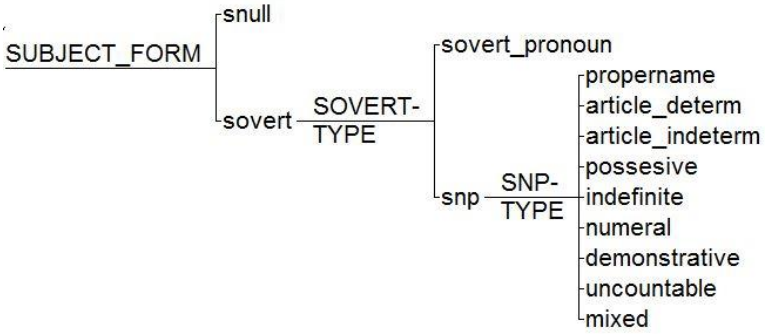


Figure 22. Overt subjects: overt pronouns merged with NPs

We acknowledge that the merge of overt pronouns and noun phrases under one and only 'overt subject' category precludes a more fine-grained analysis between these two overt anaphoric forms (in the same way that the merge of all different types of noun phrases precludes a more fine-grained analysis of each NP type[122]). There are, however, several reasons that justify this methodological decision. First, the consideration of a dependent variable with more than two categories would require the design of a multinomial statistical model and strict collaboration with a statistician (Gudmestad et al., 2013:376). Second, and related to the first, this study mainly focuses on acquisitional issues and does not aim to provide an extensive multifactorial variationist account on anaphora in Spanish L1. Third, the valid distinction between less specified (null) and more specified forms (overt: pronouns and NPs) becomes even more pronounced when overt subjects are merged. Finally, the overt forms (pronouns and NPs) have been merged in other previous studies on anaphoric subjects in Spanish as well (e.g. Montrul, 2004; Montrul & Rodríguez Louro, 2006). Consequently, the resulting category (under the label 'sovert') will be compared to the null subjects in correlation with the factors that were considered during the annotation (see section 5.5 for details on each factor). Thus, the relevant question that shall be treated in this section is to what extent several discourse factors correlate with the production of more or less specified referential forms (null vs overt subjects). The following factors will be examined: Clause Type, Switch Reference, Antecedent Form, Antecedent Distance, Antecedent Syntactic Function, Protagonisthood, New Paragraph, Active Referents and Shared Knowledge.

[122] "NP's may be classified into different types and sizes which range from descriptions (e.g. the player) to names (first and last names) carrying differential degrees of informativity" (Rasekh, 1997:2).

6.2.1 Clause Type

The three clause types (main, coordinate and subordinate clauses) were tested for potential correlation with the production of more or less specified subject forms. The results are presented in Table 12 (the original UAM CorpusTool raw frequencies can be seen in Figure 55, Figure 56 and Figure 57 in the Appendix) and reveal, overall, important differences in the production of null and overt subjects between the three clause types ($\chi2$=65.02, p<.0001):

Clause Type \ subject form		null	overt (pronouns+NPs)	sum
Main	observed count	72 **(40.91%)**	104 **(59.09%)**	176
	expected count	*105.31*	*70.69*	
	$\chi2$ value	(10.54)	(15.7)	
Coordinate	observed count	73 **(93.59%)**	5 **(6.41%)**	78
	expected count	*46.67*	*31.33*	
	$\chi2$ value	(14.85)	(22.13)	
Subordinate	observed count	74 **(66.07%)**	38 **(33.93%)**	112
	expected count	*67.02*	*44.98*	
	$\chi2$ value	(0.73)	(1.08)	
sum		219	147	366
χ^2 = 65.024, df = 2, χ^2/df = 32.51 , P(χ^2 > 65.024) = 0.0000				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 12. Clause type factor in Spanish L1

Three post-hoc pairwise comparisons were performed (alpha level adjusted at 1.8%) and revealed that Spanish native speakers produce significantly more null subjects in coordinate than in main ($\chi2$=59.10, p<.0001) and subordinate clauses ($\chi2$=18.34, p<.0001). Additionally, there are significantly more null subjects in subordinate than in main clauses ($\chi2$=16.35, p<.0001). It should be noted that this is the first study that empirically examines the influence of clause type (including coordination) in the production of anaphoric subjects (including NPs) in Spanish L1. The results reveal that it is a very important factor that affects the referential choices of native speakers, insofar as more null subjects are produced according to clause type in this order: coordinate clauses (93.59%) > subordinate clauses (66.07%) > main clauses (40.91%).

6.2.2 Switch Reference

The production of null and overt subject forms was examined in relation to same and switch reference contexts. The results are presented in Table 13 (the original UAM CorpusTool raw frequencies can be seen in Figure 58 and Figure 59 in the Appendix) and highlight the importance of this discourse factor in anaphora:

subject form / Switch Reference		**null**	**overt (pronouns+NPs)**	sum
Same-reference	observed count	157 **(86.26%)**	25 **(13.74%)**	182
	expected count	*108.9*	*73.1*	
	$\chi 2$ value	(21.24)	(31.65)	
Switch-reference	observed count	62 **(33.70%)**	122 **(66.30%)**	184
	expected count	*110.1*	*73.9*	
	$\chi 2$ value	(21.01)	(31.3)	
sum		219	147	366
χ^2 = 105.209, df = 1, χ^2/df = 105.21 , P(χ^2 > 105.209) = 0.0000				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 13. Switch Reference factor in Spanish L1

The results reveal that significantly more null subjects (88.26%) are produced in same reference than in switch reference (33.70%) contexts ($\chi 2$=105.209, p<.0001). Recall here that the relevance of Switch Reference has been consistently highlighted in numerous anaphora studies (see 5.5.2.1 for details). This finding is, thus, totally in line with the previous literature on the matter (Cameron, 1994; Geeslin & Gudmestad, 2016; Shin & Cairns, 2012; Silva Corvalán, 1982; *inter alia*).

6.2.3 Antecedent Form: priming effect

In line with the procedure followed for the forms of the anaphor, the different forms of the antecedent were also merged into two categories: null antecedents and overt antecedents. The former concerns null antecedent forms whereas the latter includes all the types of overtly expressed antecedent forms (pronouns, noun phrases, clitics etc.). Hence, the relevant question broadly concerns priming, insofar as it shall be examined whether the production of null anaphors correlates with the presence of null antecedents (and whether overt forms are triggered by the presence of overt antecedents respectively). The results are presented in Table 14 (the original UAM CorpusTool raw frequencies can be seen in Figure 60 and Figure 61 in the Appendix):

subject form / Antecedent Form		null	overt (pronouns+NPs)	sum
Null	observed count	99 **(69.72%)**	43 **(30.28%)**	142
	expected count	*84.97*	*57.03*	
	$\chi 2$ value	(2.32)	(3.45)	
Overt	observed count	120 **(53.57%)**	104 **(46.43%)**	224
	expected count	*134.03*	*89.97*	
	$\chi 2$ value	(1.47)	(2.19)	
sum		219	147	366
χ^2 = 9.428, df = 1, χ^2/df = 9.43 , P(χ^2 > 9.428) = 0.0021				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 14. Antecedent Form factor in Spanish L1

The results reveal some priming effect, in line with previous anaphora studies (see 5.5.2.2 for details). As can be seen, significantly more null subjects (69.72%) are produced when the antecedent is also null ($\chi 2$=9.42, p=.0021). Note, however, that the priming effect is not as strong as the effect of the two other previously-examined factors (Clause Type and Switch Reference) where the differences were found to be extremely significant. Additionally, regarding priming and overt subjects, we observe that the presence of overt antecedents does not seem to trigger the production of overt anaphors (more than half of the overt antecedents (53.57%) are followed by a null anaphor). Whatever the case may be, it is out of the scope of this study to provide a more extensive account on this matter.

6.2.4 Antecedent Distance

The importance of distance has been extensively highlighted in previous anaphora studies (see 5.5.2.3 for details). Due to the ordinal nature of the Antecedent Distance factor (measured in number of clauses), we start by examining how it linearly relates with the production of anaphoric subjects. In Figure 23, the frequency of anaphors according to Antecedent Distance is graphically represented (the original UAM CorpusTool raw frequencies can be seen in Figure 62 in the Appendix):

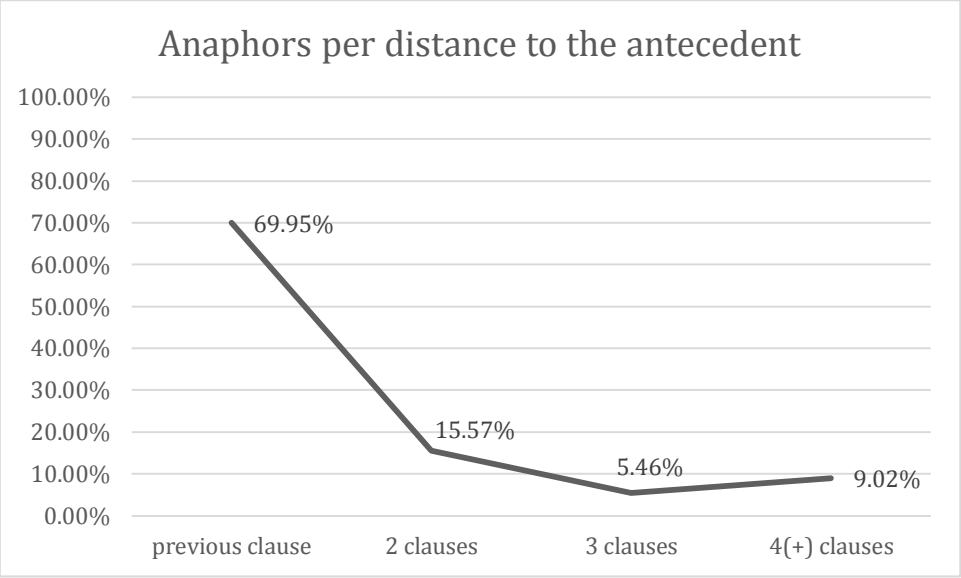


Figure 23. Antecedent Distance factor in Spanish L1 (overall distribution)

As can be observed, in the majority of cases (70%) the antecedent is located in the previous clause. Overall, there is a linear effect of distance on the production of anaphoric expressions: as distance grows, less anaphoric subjects are produced[123]. In line with the procedure followed in the previous sections, it was further examined whether Antecedent Distance correlates with the production of more or less specified anaphoric forms (null vs overt subjects). The results are presented in Table 15 (the original UAM CorpusTool raw frequencies can be seen in Figure 63, Figure 64, Figure 65 and Figure 66 in the Appendix) and reveal an overall significant correlation between Antecedent Distance and anaphoric expression ($\chi 2=89.77$, $p<.0001$):

[123] At first glance, this linearity seems to be violated by the '4(+) clauses' category. Recall, however, that the former includes all anaphoric subjects whose antecedent is located 4 or more clauses away, i.e. 5 clauses, 6 clauses, 7 clauses etc. It is reasonable to assume that, if divided, the percentage of anaphors that individually corresponds to each subcategory will be much lower.

subject form / Antecedent Distance		null	overt (pronouns+NPs)	sum
Previous clause	observed count	**191 (74.61%)**	**65 (25.39%)**	256
	expected count	*153.18*	*102.82*	
	$\chi 2$ value	(9.34)	(13.91)	
2 clauses	observed count	**23 (40.35%)**	**34 (59.65%)**	57
	expected count	*34.11*	*22.89*	
	$\chi 2$ value	(3.62)	(5.39)	
3 clauses	observed count	**4 (20%)**	**16 (80%)**	20
	expected count	*11.97*	*8.03*	
	$\chi 2$ value	(5.3)	(7.9)	
4(+) clauses	observed count	**1 (3.03%)**	**32 (96.97%)**	33
	expected count	*19.75*	*13.25*	
	$\chi 2$ value	(17.8)	(26.51)	
sum		219	147	366
χ^2 = 89.770, df = 3, χ^2/df = 29.92 , P(χ^2 > 89.770) = 0.0000				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 15. Antecedent Distance factor in Spanish L1

Six post-hoc pairwise comparisons were performed (Fisher's exact tests) in order to specifically account for the differences (alpha level adjusted at 0.8%). It was revealed that significantly more null than overt subjects were produced in the previous clause than in two, three or four (+) clause distances ($p<.0001$ for all three comparisons). On the other side, significantly more overt subjects were produced when the antecedent was four (+) clauses away than when it was two clauses away ($p=.0001$). No significant difference was found between the distances of two and three clauses ($p=.1132$), nor between three and four (+) clauses ($p=.0611$). The results, thus, confirm the important effect of Antecedent Distance on the production of anaphoric subject expressions. Although the differences are more pronounced at the edges of the distance scale, it may be concluded that overall less specified forms (null subjects) are produced more frequently when the antecedent is located closer to the anaphor. This is in line with the bulk of the previous literature on discourse anaphora (Ariel, 1990; Givón, 1983; Kibrik, 2011; *inter alia*).

6.2.5 Antecedent Syntactic Function

The relevance of the syntactic function of the antecedent for the production of anaphoric subjects has been extensively highlighted in previous anaphora studies (see 5.5.2.4 for details). More specifically, subject antecedents have been claimed to trigger the

production of null anaphors to a greater extent than non-subject antecedents. In Table 16 we see the results of the analysis regarding the syntactic function of the antecedent (the original UAM CorpusTool raw frequencies can be seen in Figure 67 and Figure 68 in the Appendix):

subject form / Antecedent Function		**null**	**overt (pronouns+NPs)**	sum
Subject	observed count	190 **(65.29%)**	101 **(34.71%)**	291
	expected count	*174.12*	*116.88*	
	$\chi 2$ value	(1.45)	(2.16)	
Non-subject	observed count	29 **(38.67%)**	46 **(61.33%)**	75
	expected count	*44.88*	*30.12*	
	$\chi 2$ value	(5.62)	(8.37)	
sum		219	147	366
χ^2 = 17.590, df = 1, χ^2/df = 17.59 , P(χ^2 > 17.590) = 0.0000				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 16. Antecedent Syntactic Function factor in Spanish L1

The results demonstrate that significantly more null subjects (65.29%) are produced when the antecedent is in subject than in non-subject position ($\chi 2$=17.59, p<.0001). This correlation is broadly in line with the findings in previous literature, especially regarding the PAS structure (see 5.5.2.5). Note, however, that the PAS only accounts for ambiguous anaphors in presence of two same-gender potential antecedents located specifically in the previous clause. The above finding, in contrast, concerns real discourse production and all kinds of anaphoric patterns (e.g. all distances, more/less than two potential antecedents, all clause types, etc.).

6.2.6 Protagonisthood

The Protagonist status of the anaphor was specifically considered during the annotation (as described in section 5.5.2.6). In this section it shall be examined whether this status is associated with the production of less specific anaphoric subjects or not. The results of the analysis are presented in Table 17 (the original UAM CorpusTool raw frequencies can be seen in Figure 69 and Figure 70 in the Appendix):

Protagonisthood \ subject form		null	overt (pronouns+NPs)	sum
Protagonist	observed count	107 **(60.80%)**	69 **(39.20%)**	176
	expected count	*105.31*	*70.69*	
	$\chi 2$ value	(0.03)	(0.04)	
Non-protagonist	observed count	112 **(58.95%)**	78 **(41.05%)**	190
	expected count	*113.69*	*76.31*	
	$\chi 2$ value	(0.03)	(0.04)	
sum		219	147	366
$\chi^2 = 0.130$, df = 1, $\chi^2/df = 0.13$, $P(\chi^2 > 0.130) = 0.7186$				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 17. Protagonisthood factor in Spanish L1

The results indicate that there is no significant association between the Protagonist status of the referent and the overall production of null or overt anaphoric forms ($\chi 2$=0.13, p=.7186). In other words, there is not an overall preference for less specified forms (null subjects) in order to refer to the protagonist of the story (as opposed to the other discourse entities). It should be noted that this is the first study that empirically examines the Protagonisthood factor (also known as Discourse Topic) in Spanish L1.

6.2.7 New Paragraph

Previous literature on anaphora has considered the starting of a new paragraph as a barrier to the production of less specified anaphoric forms (see 5.5.2.7 for details). The present study, following Lozano (2016:266) who argues for the need of research on this matter, is the first to empirically test whether the above observation holds for the production of anaphoric subjects in Spanish L1. In line with the procedure followed in the previous sections, the association of New Paragraph with the production of null and overt subjects was statistically examined. The results are presented in Table 18 (the original UAM CorpusTool raw frequencies can be seen in Figure 71 and Figure 72 in the Appendix):

subject form / New Paragraph		null	overt (pronouns+NPs)	sum
Same paragraph	observed count	211 **(65.12%)**	113 **(34.88%)**	324
	expected count	*193.87*	*130.13*	
	$\chi 2$ value	(1.51)	(2.26)	
New paragraph	observed count	8 **(19.05%)**	34 **(80.95%)**	42
	expected count	*25.13*	*16.87*	
	$\chi 2$ value	(11.68)	(17.4)	
sum		219	147	366
χ^2 = 32.844, df = 1, χ^2/df = 32.84 , P(χ^2 > 32.844) = 0.0000				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 18. New Paragraph factor in Spanish L1

The results indicate a very strong correlation between the production of more specified forms and the beginning of a new paragraph, insofar as significantly more overt (80.95%) than null forms are produced in new paragraph clauses ($\chi 2$=32.84, p<.0001). This is totally in line with the claims made in the literature on this matter (Hinds, 1977; Hofmann, 1989; Lozano, 2016) and confirms the relevance of the New Paragraph factor in Spanish L1 anaphoric discourse.

6.2.8 Active Referents

The number and gender of potentially interfering referents has been particularly considered during the annotation (see 5.5.2.8 for details). The Active Referents factor, also known as PRI (potential referential interference), aims to test whether more specified subject forms are produced in the presence of more active referents. Additionally, it shall be tested whether the gender of the potentially interfering referents is related to the production of anaphoric subjects. The results of the analysis regarding the production of null and overt subject forms for the four categories of the Active Referents factor (one, two, three or more active referents) are presented below (the original UAM CorpusTool raw frequencies can be seen in Figure 73, Figure 74, Figure 75 and Figure 76 in the Appendix):

Active Referents \ subject form		null	overt (pronouns+NPs)	sum
One	observed count expected count χ2 value	151 **(73.30%)** *123.26* (6.24)	55 **(26.70%)** *82.74* (9.3)	206
Two	observed count expected count χ2 value	53 **(51.96%)** *61.03* (1.06)	49 **(48.04%)** *40.97* (1.58)	102
Three	observed count expected count χ2 value	15 **(37.50%)** *23.93* (3.34)	25 **(62.50%)** *16.07* (4.97)	40
Four+	observed count expected count χ2 value	0 **(0%)** *10.77* (10.77)	18 **(100%)** *7.23* (16.05)	18
sum		219	147	366
χ^2 = 53.293, df = 3, χ^2/df = 17.76 , P(χ^2 > 53.293) = 0.0000				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 19. Active Referents factor in Spanish L1

The results indicate, overall, a strong association between the Active Referents factor and the production of null and overt subject forms (χ2=53.29, p<.0001). Six post-hoc pairwise comparisons with Fisher's exact tests were performed (alpha level adjusted at 0.8%) in order to specifically account for the differences. It was revealed that significantly more null forms are produced when there is only one than with two, three or more active referents (p<.0001 for all three comparisons). In addition, significantly more overt subjects are produced with four (or more) active referents than with two (p<.0001) or three (p=.0024). No significant difference was found between the anaphoric subjects with two and three potential antecedents (p=.1379). The results are in line with the previous literature, insofar as more specified subject forms are produced as the number of potential antecedents grows (Ariel, 1990; Givón, 1983; Lozano, 2016). Note, however, that the differences are more pronounced at the edges of the scale whereas no significant difference was found between the discourse patterns with two and three active referents.

Additionally, in order to test for a potential gender effect in the referential choices of the native speakers, all categories of the Active Referents factor were further annotated for

containing or not a referent of the same gender with the anaphor[124] (see 5.5.2.8 for details). The distribution of the two types (at least one same-gender referent versus no same-gender referents) in relation to the production of null and overt subject forms is presented in Table 20 (the original UAM CorpusTool raw frequencies can be seen in Figure 77 and Figure 78 in the Appendix):

subject form / Antecedent Gender		**null**	**overt (pronouns+NPs)**	sum
Same gender	observed count	39 **(43.33%)**	51 **(56.67%)**	90
	expected count	*38.25*	*51.75*	
	$\chi 2$ value	(0.01)	(0.01)	
Different gender	observed count	29 **(41.43%)**	41 **(58.57%)**	70
	expected count	*29.75*	*40.25*	
	$\chi 2$ value	(0.02)	(0.01)	
sum		68	92	160
$\chi^2 = 0.058$, df = 1, $\chi^2/df = 0.06$, $P(\chi^2 > 0.058) = 0.8092$				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 20. Antecedent Gender factor in Spanish L1

No significant differences were found regarding the production of null and overt forms in relation to the gender of the antecedents ($\chi 2$=0.06, p=.8092). Recall, however, that the overt forms contain both pronouns and noun phrases. In order to directly compare with two previous studies that have specifically examined the gender effect (Arnold & Griffin, 2007; Lozano, 2016), the two types of overt subjects were further separated and examined. The results, focusing exclusively on the overt subject forms of the previous data, are presented in Table 21 (the original UAM CorpusTool raw frequencies can be seen in Figure 79 and Figure 80 in the Appendix):

[124] Obviously, this distinction is not relevant for cases with only one potential antecedent. The corresponding analysis, thus, exclusively concerns the data of the other three types (two, three or more active referents).

overt subject form / Antecedent Gender		**overt pronoun**	**NP**	sum
Same gender	observed count	6 **(11.76%)**	45 **(88.24%)**	51
	expected count	*11.09*	*39.91*	
	$\chi 2$ value	(2.33)	(0.65)	
Different gender	observed count	14 **(34.15%)**	27 **(65.85%)**	41
	expected count	*8.91*	*32.09*	
	$\chi 2$ value	(2.9)	(0.81)	
sum		20	72	92
$\chi^2 = 6.692$, df = 1, $\chi^2/df = 6.69$, $P(\chi^2 > 6.692) = 0.0097$				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 21. Antecedent Gender factor in Spanish L1 (only overt forms)

The analysis revealed that in presence of another referent of the same gender with the anaphor, native speakers produce significantly more noun phrases than overt pronouns ($\chi 2$=6.69, p=.0097). This is in line with Lozano (2016) who also found that native speakers of Spanish prefer an NP to an overt pronoun in presence of two or more same-gender antecedents. On the other hand, this result runs against Arnold & Griffin (2007) who found that the production of more specified forms depends on the quantity but, crucially, not on the gender of the antecedents. Note, however, that the authors of the aforementioned study examined anaphora in English L1 (for the differences between English and Spanish regarding anaphoric distribution see section 3.1).

6.2.9 Shared Knowledge

The annotated anaphoric subjects were specifically tagged for the presence of Shared Knowledge constraints, as described in section 5.5.2.9. Our aim was to examine whether the overall production of less or more specific anaphoric subjects is associated with the presence of shared knowledge. The results of the analysis regarding Shared Knowledge are presented in Table 22 (the original UAM CorpusTool raw frequencies can be seen in Figure 81 and Figure 82 in the Appendix):

subject form / Shared Knowledge		null	overt (pronouns+NPs)	sum
No	observed count	175 **(58.72%)**	123 **(41.28%)**	298
	expected count	*178.31*	*119.69*	
	$\chi 2$ value	(0.06)	(0.09)	
Yes	observed count	44 **(64.71%)**	24 **(35.29%)**	68
	expected count	*40.69*	*27.31*	
	$\chi 2$ value	(0.27)	(0.4)	
sum		219	147	366
χ^2 = 0.824, df = 1, χ^2/df = 0.82 , P(χ^2 > 0.824) = 0.3640				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 22. Shared Knowledge factor in Spanish L1

The results reveal that there is no significant association between the presence of Shared Knowledge constraints and the production of null or overt subject forms ($\chi 2$=0.82, p=.3640). This finding indicates that the role of context applies to the same degree to the production of null and overt anaphoric subjects. The presence of previous discourse and/or world knowledge constraints (the two types of Shared Knowledge) does not seem to entail the selection of less specified anaphoric forms.

6.2.10 Summary of anaphora factors in Spanish L1

In sum, this section provides some answers to the research question (1) in Chapter 4. Most of the factors considered during the annotation were found to have a significant effect on the production of anaphoric forms. More precisely, less specified anaphors (null subjects) are mostly produced in coordinate clauses, in same reference contexts, when the antecedent is in the previous clause, when it is in subject position and when there is only one active referent. On the other hand, more specified forms (overt subjects) are mostly produced in main clauses, in switch reference contexts, when the antecedent is more than one clause away, in presence of more than one active referents and at the beginning of a new paragraph. Regarding the form of the antecedent, a very mild effect of priming was found: more null subjects are produced when the antecedent is also unexpressed (the same is not true, however, regarding the production of overt forms). Additionally, the gender of potential antecedents was found to affect the referential choices of the native speakers, insofar as more noun phrases than overt pronominals are produced in presence of another same-gender referent in the previous discourse. On the other hand, neither the protagonist status of the referent nor the shared knowledge constraints were found to affect the production of anaphoric subjects. Overall, the results confirm Hypothesis I insofar as the

production of anaphoric subjects in Spanish L1 was found to depend on several factors proposed in the theoretical literature. As a novelty of this study, the effect of Clause Type was examined and was found to crucially determine the production of anaphoric subjects. The results of this section confirm the complexity of the phenomenon of discourse anaphora, insofar as the production of 3rd person anaphoric subjects in Spanish L1 was found to depend on the complex interaction of several syntactic and discursive factors.

6.3 Pragmatic and unpragmatic subject expressions

In this section we turn our attention to the main focus of this study, namely the production of anaphoric subjects in L2 Spanish by English (L1) and Greek (L1) learners at several proficiency levels. A different approach than in the previous section (focusing on Spanish L1) will be adopted here. As it has been argued in section 3.3, the straightforward application of a discourse-oriented model for the exploration of L2 data may not be fine-grained enough to account for acquisitional issues. For example, if we followed the same procedure as in the previous section, we might discover that learners also produce more specific (overt) forms when starting a new paragraph or when the antecedent is far away from the anaphor. We might also find out that they do so to a lesser or greater degree than the native speakers. However, unless we specifically define pragmatic and unpragmatic (overexplicit or underexplicit) patterns and meticulously compare the L1 and L2 production of anaphoric forms in these patterns, it is difficult to make suggestions as to why L2 learners and native speakers differ (if they do). Consequently, in this section we will focus on the pragmaticality (pragmatic vs unpragmatic) instead of the specificity (overall production of null vs overt forms) of the referential choices made by native speakers and learners.

Regarding the felicity of the subject expressions, recall here that all anaphors were tagged for being pragmatic or unpragmatic according to the specific criteria described in section 5.5.3. We will start the pragmaticality analysis by examining all groups together in order to make some general observations regarding the types and distributions of unpragmatic subject forms. This will be followed by a CIA considering each proficiency level vis-à-vis to the native control group. The CIA will be complemented by an analysis of each group's interlanguage in its own right. Additionally, a developmental account will be presented by placing all English-speaking learner groups together and comparing them to each other. The same will be done for the Greek-speaking learner groups.

The overall distribution of the anaphoric subject forms (for all groups together) according to pragmaticality is displayed in Figure 24 (the original UAM CorpusTool raw frequencies can be seen in Figure 83 in the Appendix):

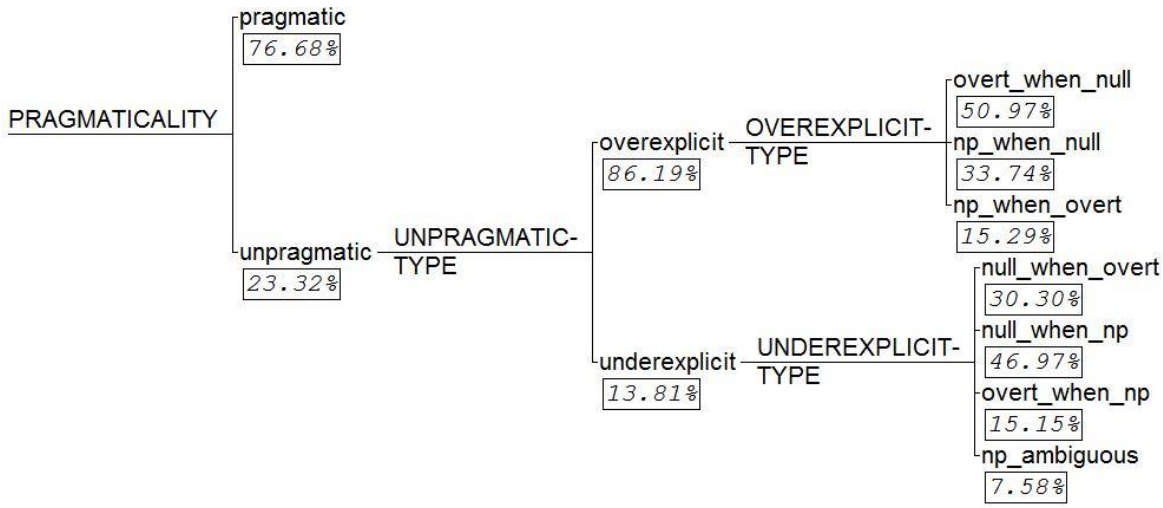


Figure 24. Pragmatic and unpragmatic anaphoric subjects in the dataset (all groups together)

In the overall distribution of the anaphoric subjects according to pragmaticality, we can observe that the large majority (76.68%) of subject expressions in the dataset are pragmatic. The unpragmatic anaphoric expressions account for the remaining 23.32% of the data. The vast majority of them (86.19%) concerns the production of overexplicit subjects, which is distributed in the three types discussed in section 5.5.3. The underexplicit subjects account for the remaining 13.81% of the unpragmatic cases. The relevant question regarding the proportions of infelicitous referential choices reported in Figure 24 concerns their distribution among the different groups. It is crucial to determine whether some group overuses or underuses pragmatic/unpragmatic forms and how this behaviour varies from one group to another, both between different proficiency levels and L1s (English, Greek, Spanish).

6.3.1 Overexplicitness and underexplicitness

Table 23 shows the proportions of pragmatic and unpragmatic anaphoric subject forms for each group by proficiency level (the original UAM CorpusTool raw frequencies can be seen in Figure 84 in the Appendix):

Group \ Felicity		Pragmatic	Overexplicit	Underexplicit	Total
Intermediate level	**English1**	169 **52.65%**	147 **45.79%**	5 **1.56%**	321 **100%**
	Greek1	168 **83.17%**	32 **15.84%**	2 **0.99%**	202 **100%**
Advanced level	**English2**	195 **69.89%**	79 **28.32%**	5 **1.79%**	279 **100%**
	Greek2	212 **86.18%**	30 **12.20%**	4 **1.62%**	246 **100%**
Upper-advanced level	**English3**	323 **76.36%**	85 **20.09%**	15 **3.55%**	423 **100%**
	Greek3	190 **89.20%**	15 **7.04%**	8 **3.76%**	213 **100%**
Native speakers		313 **85.52%**	24 **6.56%**	29 **7.92%**	366 **100%**

Table 23. Pragmatic and unpragmatic anaphoric subjects per group and proficiency level

The overexplicit and underexplicit referential choices will be examined separately, since the presence of one type during the analysis of the other could skew the results. The reason for this is that overexplicit discourse patterns, by definition, do not allow for underexplicitness (and vice versa) since a redundant form may not be ambiguous. Consider, for example, the prototypical discourse pattern (traditionally called a topic-chain) with a repeated overexplicit overt referential choice to the only existing referent:

140) María$_i$ es española. **Ella$_i$** nació en Málaga. **Ella$_i$** tiene ahora 20 años.

"Mary$_i$ is from Spain. **She$_i$** was born in Malaga. **She$_i$** is now 20 years old."

By definition, no ambiguity is possible in such a context. By extracting the underexplicit contexts from the analysis of overexplicitness, a crucial assumption is being made: for the remaining cases, a referential choice can be either pragmatic or overexplicit. Hence, we start by examining overexplicit choices in contrast to the pragmatic ones per proficiency level, excluding the underexplicit forms from the counts. Afterwards, we proceed in the same way for the underexplicit subjects.

6.3.1.1 Overexplicitness in the intermediate learners

Focusing on overexplicit subjects in the intermediate learners and after excluding the underexplicit forms, the results for each group (English1, Greek1 and native speakers) are graphically represented in Figure 25 (the original UAM CorpusTool raw frequencies can be seen in Figure 85 in the Appendix) and reveal important differences between groups. Both learner groups differ from the native speakers and from each other. English1 group produces significantly more redundant subjects than the Greek1 ($\chi2=49.01$, $p<.0001$) and the natives ($\chi2=128.92$, $p<.0001$). Greek1 group is also significantly more overexplicit than the natives ($\chi2=9.66$, $p<.001$). Interestingly, native speakers produce some overexplicit subjects as well (7.12%), a fact that previous research has also demonstrated. The groups are classified according to redundancy in this order: English1 (46.52%) > Greek1 (16%) > Natives (7.12%).

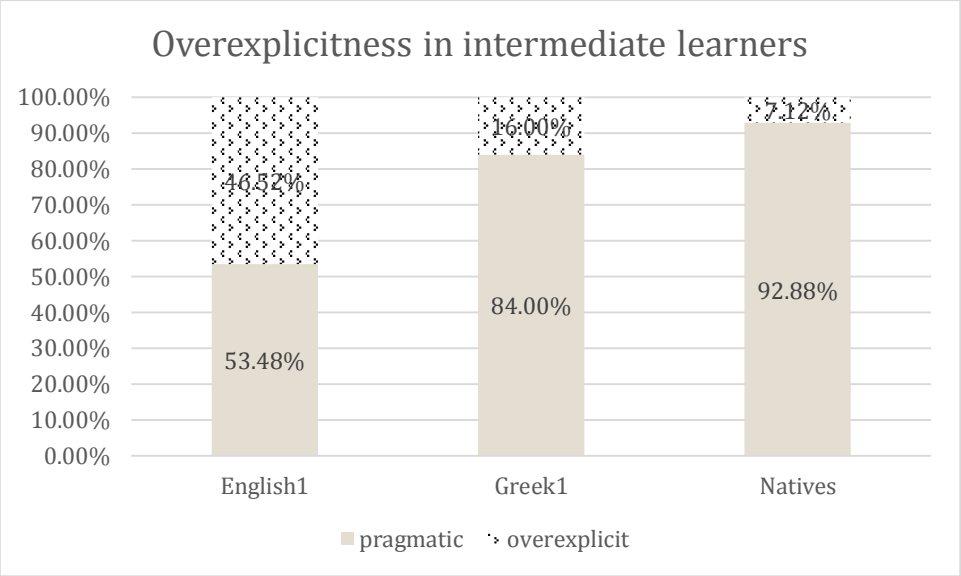


Figure 25. Intermediate learners: overexplicitness

Regarding the English1 group, nearly half of the produced anaphoric subjects are overexplicit (46.52%). The following discourse passage is representative of the referential choices made by the intermediate English-speaking group:

141) Jessica$_i$ teine pelo rubio y ojos azules. **Ella$_i$** es alta y delgada y muy bontia. Me gusta la musica de Jessica$_i$ Simpson porque es muy popular y me gusta bailo a la musica de ella$_i$. **Jessica$_i$** empezó a cantar caundo Ø$_i$ era jovenes. **Jessica$_i$** ha cantado por veinte anos. **Ella$_i$** empezó cantar a la musica Christian. Despues **ella$_i$** canta a la musica pop (ENG24_19_6_2_SH)

"**Jessica$_i$** is blonde with blue eyes. **She$_i$** is tall and thin and very beautiful. I like the music of Jessica$_i$ Simpson because it is very popular and I like dancing with her$_i$ music. **Jessica$_i$** started singing when (she)$_i$ was young. **Jessica$_i$** has sung for twenty years. **She$_i$** started singing Christian music. Then **she$_i$** sings pop music"

In example (141), although reference is being maintained constant to the one and only active referent of the discourse excerpt, only one out of seven subjects is null. The same pattern is observed, though to a lesser degree (16%), in the discourse production of the intermediate Greek-speaking group. Consider the following example extracted from this group:

142) José Antonio Domínguez Banderas$_i$ es un actor, cantante y productor de cine español. **El$_i$** nació el 10 de agosto de 1960 en una ciudad pequeña en Málaga. **Él$_i$** tiene 54 años (GR21_22_1_2_JUA)

"Jose Antonio Dominguez Banderas$_i$ is an actor, Singer and producer of Spanish movies. **He$_i$** was born at 10 of August of 1960 in a small town in Malaga. **He$_i$** is 54 years old"

Crucially, the overwhelming production of redundant overt subjects in the English1 group could be accounted for by the influence of English. Recall here that null subjects in English are, in the vast majority of discourse patterns, ungrammatical (see section 3.1). Greek, on the other hand, is a pro-drop language like Spanish. Some positive L1 influence may be at play here, since the Greek1 group may be taking advantage of the similarity between native and target language with respect to anaphoric distribution and, for that reason, performs better than the English-speaking group. However, if the only relevant factor that determines referential choices was the L1, we would expect intermediate Greek learners to be completely native-like, which is not the case. Recall here that almost all Greek-speaking learners have studied English before Spanish (see section 5.2). Some L3 influence might be also at play here. Additionally, as has been argued by some authors (Pladevall-Ballester, 2013; Polio, 1995; Rothman, 2009) learners may receive emphatic input from instructors and native speakers (more overt subjects than necessary) in the early stages of acquisition. This may explain the overexplicit production of the Greek-speaking learners and may act as a confounding factor (together with the L1 influence) in the case of the English-speaking learners. Processing factors could also account for the overexplicit production of learners from both groups (Sorace & Filiaci, 2006; Sorace, Serratrice, Filiaci, & Baldo, 2009). Finally, a universal 'ambiguity avoidance' strategy may be at play (Hendriks, 2003; Lozano, 2016; Shin & Cairns, 2009, 2012; *inter alia*). These findings require, thus, further examination and we will come back to this matter after the presentation of the results of the more proficient groups. Another finding that needs to be further explored regards the overexplicit subjects in the production of the native speakers. Although previous research has reported similar tendencies (Alonso-Ovalle, Fernández-Solera, Frazier, & Clifton, 2002; Keating, VanPatten, & Jegerski,

2011; Lozano, 2009b, 2016), a more detailed analysis is needed in order to describe the nature of the native group's overexplicitness.

With the purpose of accounting for the above findings in a more detailed manner we shall consider the overexplicit subjects produced by each group regarding the type of overexplicitness. Recall that overexplicitness was further divided into three types according to the form of the anaphoric expression (see section 5.5.3). Figure 26 shows the distribution of overexplicit subjects by type for the three groups (the original UAM CorpusTool raw frequencies can be seen in Figure 85 in the Appendix):

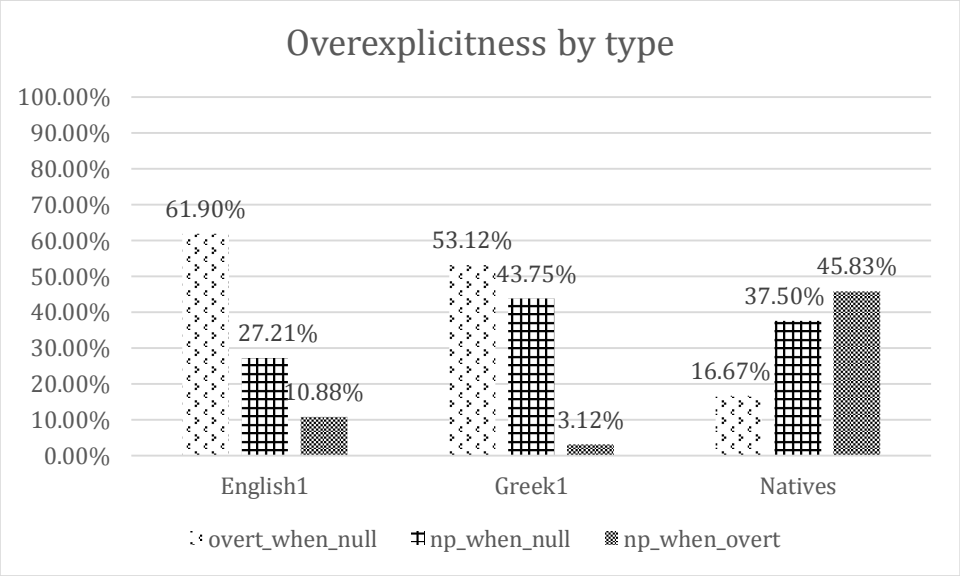


Figure 26. Intermediate learners: overexplicitness by type

As can be observed, both learner groups show the same distribution of overexplicit forms: English-speaking and Greek-speaking learners mostly overproduce overt pronouns (61.90% and 53.12% respectively) and to a lesser extent noun phrases (27.21% and 43.75% respectively), in contexts where a null subject would be appropriate. Consider the following example from the data of the intermediate English1 group:

unpragmatic/overexplicit/overt_when_null

143) El$_i$ es el actor principal en muchas peliculas que muy popular en muchos paises del mundo. **El$_i$** es americano y sus peliculas son americanos pero existen en otros paises tambien. **El$_i$** es muy guapo (ENG23_20_5_2_RW)

"He$_i$ is the main actor in many very popular films in many countries in the world. **He$_i$** is American and his films are American but they also exist in other countries. **He$_i$** is very handsome"

Crucially, there are no significant differences between the two learner groups (English1 vs Greek1) for any type of overexplicit reference ('overt_when_null': $\chi 2=0.47$, p=.4889,

'np_when_null': $\chi 2=2.67$, p=.1021, 'np_when_overt' (Fisher's exact test): p=.2095). On the other hand, both learner groups produce significantly more overt pronouns ('overt_when_null' type) than the native speakers (for English1: $\chi 2=15.31$, $p<0.001$ and for Greek1: $\chi 2=6.30$, p=.0121). Native speakers, instead, mostly overproduce noun phrases when an overt pronoun would be sufficiently informative (45.83%). Regarding the 'np_when_overt' category they differ significantly from both learner groups (for English1: $\chi 2=12.86$, p<.001 and for Greek1: $\chi 2=10.58$, p=.0011). There are no significant differences between the three groups regarding the 'np_when_null' overexplicitness type (between Greek1 and English1: $\chi 2=2.67$, p=.1022, between natives and English1: $\chi 2=0.62$, p=.4310, between natives and Greek1: $\chi 2=0.04$, p=.8414).

It should be noted here that there is an important qualitative difference between the three types of overexplicitness. When the overexplicit reference concerns the use of a noun phrase instead of a null subject or an overt pronoun, there might be some stylistic reasons involved in the referential choice. This analysis follows Lozano (2016) who was the first to consider the 'np_when_overt' type under the label 'uneconomical'. Crucially, 'uneconomical' subjects are, as the label indicates, less redundant than the purely overexplicit ones. 'Uneconomical' subjects may even stylistically enrich discourse in some cases. Consider the following example from the native data:

unpragmatic/overexplicit/np_when_overt

144) La trama trata sobre un policía$_i$ californiano que recibe un día una carta de una exnovia$_j$ a la que Ø$_i$ todavía no ha terminado de olvidar. **La chica$_j$** le$_i$ había dejado plantado en el altar (ESP24_3_JAO)

"The script is about a Californian póliceman$_i$ who receives a letter from an ex-girlfriend$_j$ that he$_i$ has not yet forgotten. **The girl$_j$** had left him$_i$ at the altar"

Essentially, in example (144), the overexplicit anaphoric noun phrase ("the girl") is not identical but synonymous to the antecedent noun phrase ("ex-girlfriend"). In this kind of overexplicitness, although an overt pronoun would be sufficiently informative, a noun phrase may be used instead without any negative pragmatic effects. On the contrary, this might entail a purposeful, though technically overexplicit, referential choice in order to

enrich the produced discourse[125]. Future research should specifically focus on this fine-grained stylistic phenomenon. On the other side, there is no stylistic enrichment involved in the overexplicit overt pronouns mostly produced by the intermediate learners (see examples (141), (142) and (143) from the learner data). Quite the opposite is true, since overexplicit pronominals may render the discourse highly unnatural and harder to process.

Given that the overexplicit choices in the native Spanish discourse seem to be, at least to some extent, stylistic and that the present investigation primarily focuses on deficits (in terms of differences with the native speakers) in the interlanguage of English and Greek learners of Spanish, we shall now independently examine the overexplicit subjects in the production of the two learner groups. The relevance of several factors that have been previously reported to account for deficits in anaphoric production will be examined, namely: number, animacy, clause type and PRI. Furthermore, the PAS contexts will be analysed separately. During the analysis, we first examine each learner group's production on its own and then compare the learner groups to each other.

GRAMMATICAL NUMBER

Starting with grammatical number, Figure 27 shows the frequency of overexplicit choices for singular and plural subjects for the two intermediate learner groups (the original UAM CorpusTool raw frequencies can be seen in Figure 86 and Figure 87 in the Appendix):

[125] The repetition of noun phrases (but, crucially not of overt pronouns) in written discourse is often due to stylistic choices. As Emmott (1997:442) points out: "For special stylistic uses, a writer may sometimes increase the lexical density in a reference chain to achieve special effects".

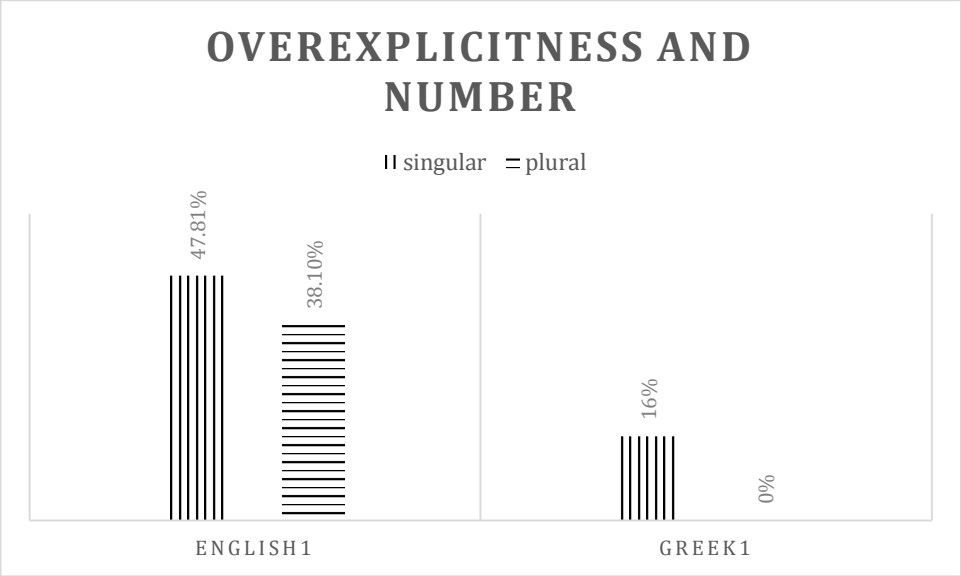


Figure 27. Intermediate learners: overexplicitness and grammatical number

We observe that for both groups, 3rd person singular referents are generally more problematic than the plural ones. Notice, however, that plural referents are not totally impervious to overexplicitness for the English1 group, as can be seen in the following example extracted from the production of a learner of this group:

145) Ella meets sus roommates Duke y sus dos amigos. Todos los ellos$_i$ jugan fútbol. **Ellos$_i$** van al practicar fútbol (ENG27_19_4_3_OM)

"She meets her roommate Duke and his two friends. All of them$_i$ play football. **They$_i$** go practice football"

The Greek1 group, which is overall less redundant, shows no overexplicit production with plural referents. Nevertheless, for both groups, the differences between frequencies of overexplicit subjects in singular and plural number are not significant (singular vs plural for the English1 group: $\chi 2=1.02$, p=.3128 and for the Greek1 group (Fisher's exact test): p=.5959). Thus, overall, the results indicate that intermediate learners are equally overexplicit with singular or plural anaphoric subjects.

ANIMACY

Another factor that has been previously found to affect anaphoric production is the animacy of the referent. Overall, in Figure 28 we see that both learner groups are less overexplicit with inanimate than with animate referents (the original UAM CorpusTool raw frequencies can be seen in Figure 88 and Figure 89 in the Appendix). However, some overexplicit production with inanimate referents has also been detected. Consider

the example below, extracted from the production of an intermediate English-speaking learner:

146) Mucha gente pensaba que Brasil$_i$ iba a ganar La Copa Mundial en Alemanía en 2006. Pero Ø$_i$ no ganó. **Brazil$_i$** perdío con Francía$_j$ 1-0. **Brazil$_i$** ganó La Copa Mundial en 2002 en Korea del sur y Japon (ENG30_29_5_2_TS)

"Many people thought that Brasil$_i$ would win the World Cup in Germany in 2006. But (it)$_i$ did not win. **Brasil$_i$** lost from France$_j$ 1-0. **Brasil$_i$** won the World Cup in 2002 in South Korea and Japan"

The differences, for both groups, between overexplicitness in animate and inanimate referents are not statistically significant (for English1: $\chi 2=3.32$, p=.0683, for Greek1 (Fisher's exact test): p=.2204). Accordingly, results indicate that intermediate learners produce overexplicit subjects with both animate and inanimate referents.

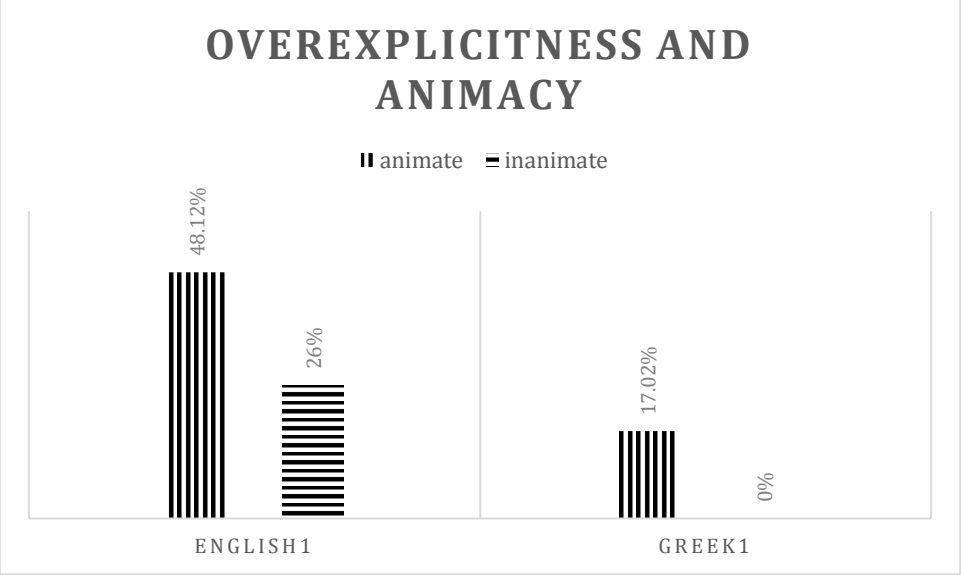


Figure 28. Intermediate learners: overexplicitness and animacy

To my knowledge, the only other study in Spanish L2 that has particularly focused on possible effects of animacy and number in unpragmatic learner production is Lozano (2009b). The author concluded that deficits are selective, since the English-speaking learners examined in his study were found to produce significantly more unpragmatic subjects with animate singular than with animate plural or inanimate singular referents. It should be noted, however, that the proficiency level of the groups examined in Lozano's study was lower-advanced and upper-advanced (proficiency test scores were minimum 90%) whereas the groups examined in this section are composed of intermediate learners (proficiency test scores are below 70%). This suggests that there might be a developmental association regarding the effect of animacy and number on

overexplicit referential choices. We shall come back to this issue after examining the more advanced groups.

CLAUSE TYPE

We shall now focus on the different clauses produced by the intermediate learners in this study in order to examine the possible influence of clause type on overexplicitness for each group. Recall here that anaphoric subjects were tagged for belonging either to a main, a coordinate or a subordinate clause. In order to compare the frequencies of overexplicit subjects in each clause type for the English1 and Greek1 groups, two independent 3x2 chi-square tests were performed.

Starting with the English1 group, in Table 24 we see the results of the statistical analysis (the original UAM CorpusTool raw frequencies can be seen in Figure 90, Figure 91 and Figure 92 in the Appendix):

felicity / clause type		**Pragmatic subjects**	**Overexplicit subjects**	sum
Main	observed count	78 **(41.93%)**	108 **(58.07%)**	186
	expected count	*99.47*	*86.53*	
	χ2 value	(4.64)	(5.33)	
Coordinate	observed count	65 **(87.83%)**	9 **(12.17%)**	74
	expected count	*39.58*	*34.42*	
	χ2 value	(16.33)	(18.78)	
Subordinate	observed count	26 **(46.42%)**	30 **(53.58%)**	56
	expected count	*29.95*	*26.05*	
	χ2 value	(0.52)	(0.6)	
sum		169	147	316
χ^2 = 46.195, df = 2, χ^2/df = 23.10 , P(χ^2 > 46.195) = 0.0000				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 24. English1 group: overexplicitness by clause type

Overall, there are highly significant differences between the production of overexplicit subjects in the three clause types (χ2=46.19, p<.0001). After performing three post-hoc pairwise comparisons (alpha level adjusted at 1.6%) we found that the English1 group employed significantly fewer overexplicit subjects in coordinate than in main (χ2=43.23, p<.0001) or in subordinate clauses (χ2=24.09, p<.0001). No differences were found between main and subordinate clauses (χ2=0.19, p=.6629). Crucially, one of the very few discourse structures in English that allow null subjects is coordination. According to

Nariyama (2004:240) "subject ellipsis in English is restricted to coordinate structures and non-finite clauses". Since non-finite clauses are, by definition, excluded from the analysis of subject forms, the only relevant structure in this study that allows the alternation of null and overt subjects in both English and Spanish is coordination. Several studies in English L1 have focused on subject ellipsis in coordinate clauses (Beavers & Sag, 2004; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Gardelle & Sorlin, 2015; Haegeman, 1997). However, this is the first study in Spanish L2 that thoroughly examines how this phenomenon may affect the interlanguage of English-speaking learners of Spanish. As we can see, more than half of the subjects produced by the English1 group in main and subordinate clauses are overexplicit (58.07% and 53.08% respectively). This percentage drops drastically to 12.17% in coordinate clauses. Note, however, that this proportion includes both same- and different-subject coordinates. The distinction between the two types is made explicit in the examples below:

147) Ella$_i$ levanta el próximo día y **ella$_i$** tiene 30 años pero su vida no esta como ella$_i$ esperaría (ENG29_20_ _2_EMH)

"She$_i$ gets up the next day and **she$_i$** is 30 years old but her life is not how she$_i$ expected"

148) Wheeler$_i$ salió Ronnie$_j$ por una chica y **Ronnie$_j$** caminado a su casa (ENG30_20_6_3_NJP)

"Wheeler$_i$ left Ronnie$_j$ for a girl and **Ronnie$_j$** went home"

Note that the only structure that allows null subjects in English is same-subject coordination, as represented in example (147). Hence, in order to perceive the full amplitude of the phenomenon we analysed separately the same-reference coordinates, by excluding the switch-reference coordinates from the data in Table 24. As a result, the percentage of overexplicit subjects further dropped to 5.55% (there are only 3 cases of overexplicit anaphors in same-subject coordination in the whole English1 dataset). The results are shown in Table 25 (the original UAM CorpusTool raw frequencies can be seen in Figure 93 in the Appendix):

clause type \ felicity		Pragmatic subjects	Overexplicit subjects	sum
Main	observed count	78 **(41.93%)**	108 **(58.07%)**	186
	expected count	*97.4*	*88.6*	
	χ2 value	3.86	4.25	
Coordinate (only same-subject)	observed count	51 **(94.45%)**	3 **(5.55%)**	54
	expected count	*28.28*	*25.72*	
	χ2 value	18.26	20.07	
Subordinate	observed count	26 **(46.42%)**	30 **(53.58%)**	56
	expected count	*29.32*	*26.68*	
	χ2 value	0.38	0.41	
sum		155	141	296
χ^2 = 47.235, df = 2, χ^2/df = 23.62 , P(χ^2 > 47.235) = 0.0000				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 25. English1 group: overexplicitness in same-subject coordination

This finding further corroborates the premise that English-speaking intermediate learners of Spanish may transfer the null subject ellipsis feature in same-subject coordination from their L1. To my knowledge, the only other study that has examined same-subject coordination as a convergent context in English and Spanish is Shin & Montes-Alcalá (2014). The authors found that second-generation native Spanish speakers who live in New York overproduce overt pronouns under the influence of English (attritional effect) but, crucially, they do so to a lesser degree in same-subject coordinate clauses. In line with this, the results of the present study indicate that intermediate proficiency English-speaking learners of Spanish L2 overproduce redundant overt pronouns, but crucially, they do so to a much lesser degree in the convergent (between Spanish and English) same-subject coordination context due to the influence of their L1.

Subsequently, the same procedure was followed for the Greek1 group. We start by comparing the frequencies of overexplicit subjects in the three clause types. For that purpose, a 3x2 chi-square test was performed. Results are presented in Table 26 (the original UAM CorpusTool raw frequencies can be seen in Figure 90, Figure 91 and Figure 92 in the Appendix):

clause type \ felicity		Pragmatic subjects	Overexplicit subjects	sum
Main	observed count	106 **(79.10%)**	28 **(20.89%)**	134
	expected count	*112.56*	*21.44*	
	$\chi 2$ value	(0.38)	(2.01)	
Coordinate	observed count	45 **(97.82%)**	1 **(2.18%)**	46
	expected count	*38.64*	*7.36*	
	$\chi 2$ value	(1.05)	(5.5)	
Subordinate	observed count	17 **(85%)**	3 **(15%)**	20
	expected count	*16.8*	*3.2*	
	$\chi 2$ value	(0)	(0.01)	
sum		168	32	200
$\chi^2 = 8.947$, df = 2, $\chi^2/df = 4.47$, $P(\chi^2 > 8.947) = 0.0114$				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 26. Greek1 group: overexplicitness by clause type

Likewise the English1 group, in the Greek1 group there is a significant overall association between clause type and overexplicitness ($\chi 2=8.94$, p=.0114). Due to the low raw frequencies (≤5 cases) of overexplicit subjects in some cells, three post-hoc pairwise comparisons with Fisher's exact tests were performed (main vs coordinate, main vs subordinate and coordinate vs subordinate). The results regarding coordination are quite similar to those of the English1 group (see Table 24): there are significantly fewer overexplicit subjects in coordinate compared to main clauses (p=.0044). However, although Greek1 group is also less overexplicit in coordinate clauses as compared to subordinate clauses, this difference approximates but does not reach significance (p=.0791). Additionally, there is no significant difference in the production of overexplicit subjects between main and subordinate clauses (p=.5793). Intermediate Greek-speaking learners of Spanish, as we have seen, seem to take advantage of their L1 properties and produce overall less overexplicit subjects than their English-speaking counterparts. Considering that coordination is a practically categorical subject-ellipsis context in Greek, it is reasonable to assume that they will produce even less overt subjects in this context. On the other hand though, the one and only case of overexplicitness in coordination structures for this group further corroborates the hypothesis that L1 is not the only relevant factor in L2 anaphoric production. Consider the following example extracted from the Greek1 group:

149) Manu$_i$ es de Espana, pero **el$_i$** nacido en Francia (GR22_21_0_2_christos)

"Manu$_i$ is from Spain, but **he$_i$** was born in France"

Although a null subject would be the pragmatic choice in the same-subject coordinate clause of the example (149) in both Greek and Spanish, the Greek-speaking learner in question chooses a redundant overt pronoun. As already argued, this may entail some L3 influence (since Greek-speaking learners of Spanish have additionally some previous knowledge of English), a universal 'ambiguity avoidance' strategy, processing factors or some input-related causes. These accounts shall be broadly considered in the summary of the section and in the general discussion.

PRI

We will now turn to the potential referential interference (PRI, also known as 'number of potential antecedents' in the literature) in order to examine how the presence of other active referents may trigger or not the presence of overexplicit anaphoric subjects. We start by examining the English1 group's production in the four categories of the Active Referents tag. Following the same statistical analysis as in the Clause Type category, a 4x2 chi-square test was performed. Results are presented in Table 27 (the original UAM CorpusTool raw frequencies can be seen in Figure 94, Figure 95, Figure 96 and Figure 97 in the Appendix):

felicity / active referents		**Pragmatic subjects**	**Overexplicit subjects**	sum
one_ref	observed count	82 **(42.26%)**	112 **(57.74%)**	194
	expected count	*103.75*	*90.25*	
	$\chi2$ value	(4.56)	(5.24)	
two_ref	observed count	49 **(63.63%)**	28 **(36.36%)**	77
	expected count	*41.18*	*35.82*	
	$\chi2$ value	(1.48)	(1.71)	
three_ref	observed count	19 **(76%)**	6 **(24%)**	25
	expected count	*13.37*	*11.63*	
	$\chi2$ value	(2.37)	(2.73)	
fourplus	observed count	19 **(95%)**	1 **(5%)**	20
	expected count	*10.7*	*9.3*	
	$\chi2$ value	(6.45)	(7.41)	
sum		169	47	316
χ^2 = 31.950, df = 3, χ^2/df = 10.65 , P(χ^2 > 31.950) = 0.0000				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 27. English1 group: overexplicitness and PRI

Overall, the results indicate that there is a significant effect of number of active referents on overexplicitness ($\chi2$=31.95, p<.0001). Due to the low number of overexplicit subjects

(<5) in the 'fourplus' category, we performed six post-hoc pairwise comparisons with Fisher's exact test in order to determine the locus of variability (alpha level adjusted at 0.8%). The tests revealed that the English1 group produces significantly more overexplicit subjects with one active referent than with two (p=.0018), three (p=.0022) and four active referents (p<.001). There is no significant difference between two and three (p=.3315), between three and four (p=.1117) or between two and four referents (p=.0118). The results, thus, indicate that intermediate English-speaking learners produce more overexplicit subjects when there is no potential referential interference. This result may seem counterintuitive, since one might expect that the presence of other active referents in the immediately preceding discourse would trigger the need to be more explicit in order to avoid any potential ambiguity. It should be reminded here, however, that the majority of overexplicit subjects produced by the English1 group concern overt pronouns is same- reference structures with only one potential antecedent (see Figure 26). No ambiguity is at issue is such contexts. Therefore, the results in Table 27 further corroborate the hypothesis that redundancy in the intermediate English group participants mostly originates from negative cross-linguistic influence from their L1.

The exact same procedure was followed for the Greek1 group. A 4x2 chi-square test was performed and the results are presented in Table 28 (the original UAM CorpusTool raw frequencies can be seen in Figure 94, Figure 95, Figure 96 and Figure 97 in the Appendix):

felicity / active referents		**Pragmatic subjects**	**Overexplicit subjects**	sum
one_ref	observed count	138 **(83.13%)**	28 **(16.87%)**	166
	expected count	*139.44*	*26.56*	
	$\chi 2$ value	(0.01)	(0.08)	
two_ref	observed count	15 **(88.23%)**	2 **(11.77%)**	17
	expected count	*14.28*	*2.72*	
	$\chi 2$ value	(0.04)	(0.19)	
three_ref	observed count	11 **(84.61%)**	2 **(15.39%)**	13
	expected count	*10.92*	*2.08*	
	$\chi 2$ value	(0)	(0)	
fourplus	observed count	4 **(100%)**	0 **(0%)**	4
	expected count	*3.36*	*0.64*	
	$\chi 2$ value	(0.12)	(0.64)	
sum		168	32	200
χ^2 = 1.085, df = 3, χ^2/df = 0.36 , P(χ^2 > 1.085) = 0.7806				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 28. Greek1 group: overexplicitness and PRI

Overall, for the Greek1 group, there are no significant differences between the four conditions ($\chi 2=1.08$, $p=.7806$). Recall here that overexplicitness in the Greek1 group, same as in English1 group, concerns primarily the use of redundant overt pronouns in same-reference contexts (see Figure 26). In line with the results for the English1 group, potential ambiguity in terms of referential interference does not seem to be the crucial factor that induces the intermediate Greek-speaking learners to employ more explicit forms than necessary. Therefore, the source of their infelicitous referential choices should be sought elsewhere (L3 influence, input, processing). We shall return to this matter in the summary of this section and in the general discussion, where all proficiency levels will be considered together.

PAS

We will finish the analysis of the intermediate learners by focusing on the PAS in learner discourse. One of the aims of this investigation was to separately examine PAS structures in the learner data in order to validly compare the results with the bulk of experimental studies. However, as we can see in Table 29, the PAS structures barely represent the 3.12% of the English1 data and only 1% of the Greek1 data (the original UAM CorpusTool raw frequencies can be seen in Figure 98 in the Appendix). Overall, the total number of cases is insufficient for the application of valid statistical comparisons regarding the PAS structures.

PAS / Group	**PAS not applicable**	**PAS Intersentential**	**PAS Intrasentential**	**Total**
English1	311	5	5	321
	96.88%	**1.56%**	**1.56%**	**100%**
Greek1	200	2	0	202
	99%	**1%**	**0%**	**100%**

Table 29. Intermediate learners: PAS structures

It should be reminded here that the PAS may account only for a very specific discourse structure which follows this pattern: a main clause with two same-gender referents (one in subject and the other in non-subject position), followed by a globally ambiguous clause whose subject corefers with one of the two aforementioned referents (the order of the clauses may be also reversed). All in all, the above anaphoric pattern is very infrequent

in the discourse of the intermediate learners of this study. Recall here that the antecedent of a subject expression in real discourse may not appear in the immediately preceding clause. It may not be neither in subject nor in object position. Additionally, there may be no other active referents or, reversely, there may be more than one (or maybe just one, but of different gender). All the above scenarios are just a handful of the numerous anaphoric patterns in real discourse that do not coincide with the PAS structures. We shall return to the role of PAS in discourse after computing and analysing the production of the more advanced groups where it may be the case that more PAS structures will appear.

SUMMARY (intermediate learners)

In sum, regarding overexplicitness in the production of anaphoric subjects by intermediate learners, some important findings may be pointed out before proceeding with the rest of the analysis. First, both groups of intermediate learners of this study present deficits by producing overexplicit anaphoric 3rd person subject expressions (mostly overt pronouns) in the same discourse patterns that native speakers prefer less explicit forms. This finding supports Hypothesis II (a), insofar as learners of both groups perform in non-target manner. It should be noted, however, that English-speaking learners of Spanish are significantly more overexplicit than their Greek-speaking counterparts. This finding supports Hypothesis V, insofar as the L1 seems to be a facilitating factor. On the other hand, native speakers are also occasionally overexplicit. In their case, though, overexplicitness is of different nature (they mostly produce redundant noun phrases). Secondly, both learner groups are less redundant when it comes to certain structures that, in their respective native languages, less explicit forms are preferred. Crucially, the English group approximates native-like performance in the production of null subjects in same-subject coordinate clauses. Cross-linguistic influence might account for this, since null subjects are also allowed in these clauses in English L1 (but crucially not in any of the other clause types examined). Regarding the facilitating role of the L1, this finding is also broadly in line with Hypothesis V. Thirdly, PRI does not affect overexplicitness in learner discourse. Both groups produce redundant pronominals, even in total absence of other potential competing referents. Whilst transfer might account for this phenomenon regarding the English group, due to the fact that overt pronouns are obligatory in English, no such explanation is valid for Greek learners whose L1 promotes the use of null subjects in such contexts. Overall, the evidence presented in this section supports Hypothesis VI, insofar as no single factor seems to account for the above findings altogether. Finally, regarding PAS structures, it was not possible to test

the pertinent hypotheses due to almost inexistent number of cases in the discourse of both the learner groups. In the next section, we turn our focus to the overexplicit production of the advanced learners.

6.3.1.2 Overexplicitness in the advanced learners

Turning our focus to the advanced learners, the results for learners of both groups (English2 and Greek2) and native speakers regarding frequencies of overexplicit subjects are graphically represented in Figure 29 (the original UAM CorpusTool raw frequencies can be seen in Figure 85 in the Appendix) and reveal important differences between groups. Both learner groups differ from the native speakers and from each other. More specifically, the English2 group produces significantly more redundant subjects than the Greek2 ($\chi2=20.26$, $p<.0001$) and the natives ($\chi2=49.29$, $p<.0001$). The Greek2 group is also more overexplicit (although marginally significant) than the natives ($\chi2=4.03$, $p=.0446$). The groups are classified according to redundant production in this order: English2 (28.83%) > Greek2 (12.40%) > Natives (7.12%).

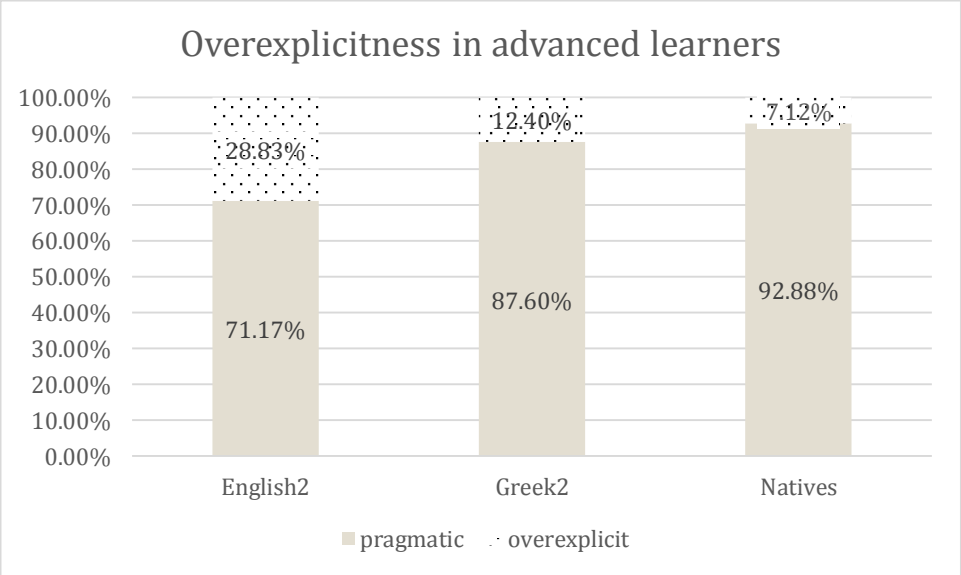


Figure 29. Advanced learners: overexplicitness

Regarding overexplicitness by group, the results for the advanced learners are similar to those of the intermediate groups (see Figure 25). Advanced English-speaking learners produce a significant amount of overexplicit subjects (28.83%), although to a lesser degree than the corresponding intermediate English1 group (46.52%). Consider the following example from the advanced English2 group:

150) La persona$_i$ que yo admiro mucho es famosa pero no es una actriz. **Ella$_i$** hizo mucho por las personas que no tuvieron mucho. **Ella$_i$** dedicó toda su vida para ayudar otras personas. **Ella$_i$** fue una misionaria (ENG33_18_4_2_MAN)

"The person$_i$ that I admire a lot is famous but it is not an actress. **She$_i$** did a lot for the people that did not have much. **She$_i$** dedicated her life to help other persons. **She$_i$** was a missionary"

In example (150) we observe the same kind of constant overexplicit referential choices that had been detected in the production of the intermediate English1 group. The same pattern, though to a much lesser frequency, is found in the discourse of the advanced Greek-speaking learners. Consider the example below extracted from the production of this group:

151) Adaline$_i$ sale de la casa porque Ø$_i$ se siente que su secreto corre peligro. **Ella$_i$** pensaba mucho y Ø$_i$ decide hacer algo (GR38_21_5_3_IFI)

"Adaline$_i$ gets out of the house because (she)$_i$ feels that her secret is in danger. **She$_i$** thinks a lot and (she)$_i$ decides to do something"

It should be noted, however, that the two learner groups of the same proficiency level (English2 and Greek2), differ regarding the frequencies in the production of overexplicit subject forms, as can be seen in Figure 29. Regarding both same-proficiency (but different L1) groups, there seems to be a parallel developmental trend, insofar as more native-like performance is achieved with higher proficiency level. Note that although the advanced English2 group presents a significant amount of unpragmatic referential production, it is not to the same degree with the intermediate English1 group. Similarly, the Greek2 advanced group performs better than the intermediate Greek1 group, approximating (but not reaching) native group's performance. The significantly better performance of the Greek2 group with respect to the English2 group could be accounted for by considering the similarity between Greek and Spanish (as opposed to the difference between English and Spanish) regarding anaphoric distribution. The reader is referred to section 6.3.1.1 for more details on the discussion of this. Regarding the differences between proficiency levels, a full developmental account will be provided on the last section of the Results.

With the purpose of accounting for the above findings in a more detailed manner we shall now consider the type of overexplicit subjects produced by each group. In line with the procedure followed for the intermediate learners, the nature of overexplicitness will be examined according to the form of the anaphoric expression (see section 5.5.3). Figure 30 shows the distribution of overexplicit subjects by type for the three groups (the original UAM CorpusTool raw frequencies can be seen in Figure 85):

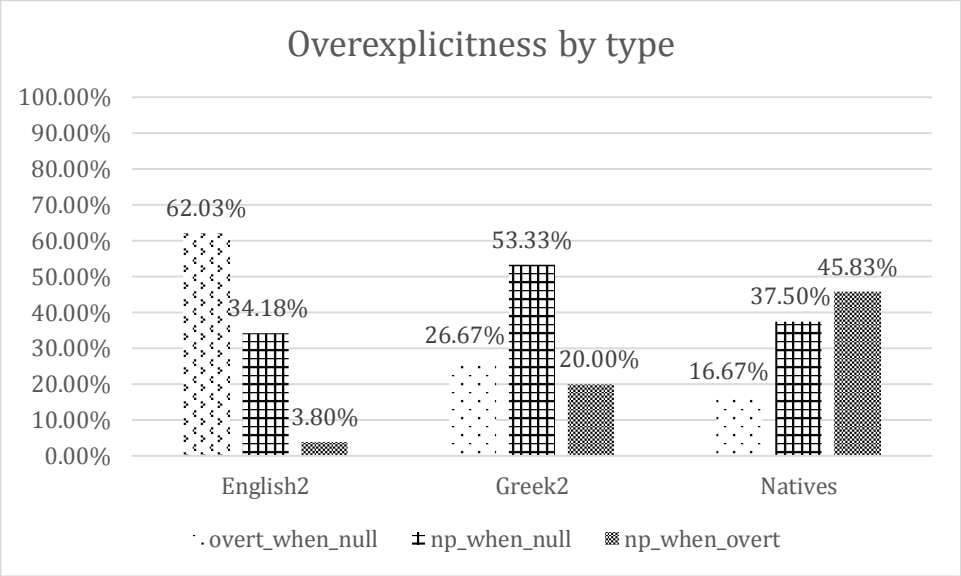


Figure 30. Advanced learners: overexplicitness by type

Regarding the English2 group, we can observe that the pattern of overexplicitness by type is similar to the one that was observed for the English1 group (see Figure 26). Advanced English-speaking learners produce an important amount of unpragmatic pronominal subjects, in cases where a null subject would be sufficiently informative. More specifically, they produce significantly more overexplicit pronouns than the Greek2 group ($\chi2$=9.53, p=.0022) and the natives ($\chi2$=13.40, p<.001). On the other hand, the advanced Greek-speaking learners do not exhibit the same behavior neither with their English counterparts nor with the corresponding Greek1 group. They mostly produce overexplicit noun phrases (53.33%), in contexts where a null subject would be pragmatically more appropriate (according to the criteria in section 5.5.3). Consider the following example from the Greek2 data:

152) Pero, ella$_i$ tiene que irse hasta las doce de medianoche, porque despues los hechizos se van. Por eso, **Cenicienta$_i$** abandona la fiesta (GR37_24_3_3_NAT)

"But she$_i$ has to go until midnight because after the spell will be gone. That´s why **Cenicienta$_i$** abandons the party"

Native speakers also, as we have already seen, mostly overproduce noun phrases (45.83%). According to the Fisher's exact tests performed, as regards the 'np_when_overt' category, they differ significantly from both advanced learner groups (for English2: p<.0001 and for Greek2: p=.0412). Crucially, as it has been previously discussed, this type of 'uneconomical' overexplicitness may entail some stylistic purposes. Though technically redundant, an overexplicit noun phrase is pragmatically

more appropriate than an overexplicit overt pronoun. Therefore, it should be considered whether the roots of this phenomenon for each type of overexplicitness might be of different nature. Notice, additionally, that the Greek2 group behaves somewhat similarly to the native group regarding overexplicitness type: overall, both groups mostly employ redundant noun phrases (in contrast to the English learners who mostly produce overexplicit pronominals). The reader is referred to the comments already made regarding intermediate learners and overexplicitness by type for these groups in section 6.3.1.1. Additionally, this matter shall be broadly considered in the general discussion.

We shall now exclusively focus on the two advanced learner groups in order to independently examine their production of overexplicit subjects in relation to other discourse factors. In line with the procedure followed for the intermediate groups, the following factors will be considered: number, animacy, clause type and PRI. Furthermore, we shall specifically focus on PAS contexts. The production of each group will be first examined separately and then compared to the production of the other group.

NUMBER

Regarding the grammatical number, Figure 31 demonstrates the frequency of overexplicit choices for singular and plural subjects for the two advanced learner groups (the original UAM CorpusTool raw frequencies can be seen in Figure 86 and Figure 87 in the Appendix):

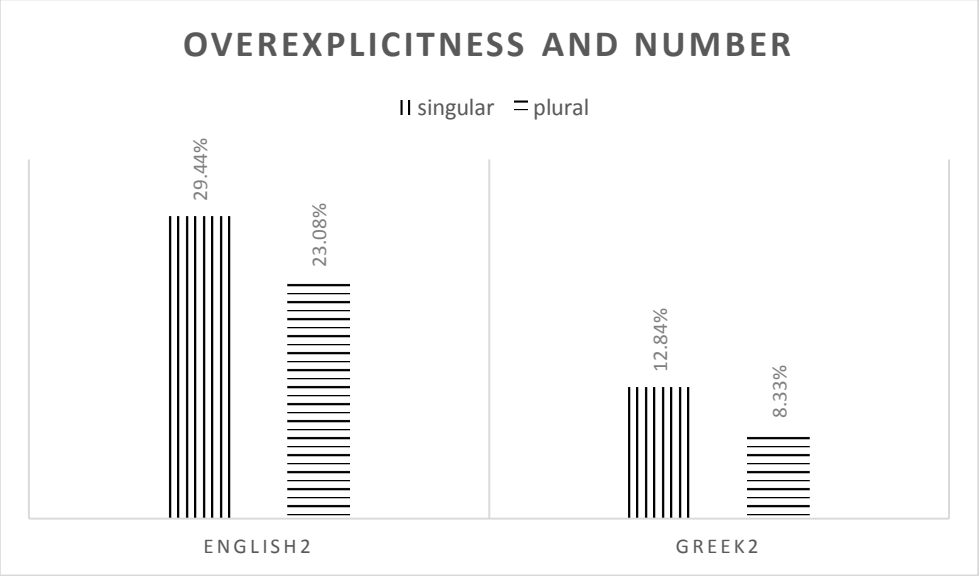


Figure 31. Advanced learners: overexplicitness and grammatical number

In line with the results of the intermediate groups, we observe that for both advanced learner groups, 3rd person singular referents are more problematic than the plural ones.

Nevertheless, for both groups, the differences between frequency of overexplicit subjects in singular and plural number are not significant (singular vs plural for English2: $\chi 2=0.21$, p=.6467 and for Greek2 (Fisher's exact test): p=.7419). We can thus conclude that, same as with the intermediate proficiency groups, the advanced learners are overexplicit with both singular and plural anaphoric subjects.

ANIMACY

Next, we shall turn our attention to the potential effect of animacy in the production of overexplicit subjects for the advanced learner groups. In Figure 32 we see that both groups are more overexplicit with animate than with inanimate referents (the original UAM CorpusTool raw frequencies can be seen in Figure 88 and Figure 89 in the Appendix). The differences, however, are not statistically significant (for English2: $\chi 2=3.31$, p=.0658 and for Greek2 (Fisher's exact test): p=.7057). Overall, results indicate that advanced learners produce overexplicit subjects with both animate and inanimate referents.

Figure 32. Advanced learners: overexplicitness and animacy

Accordingly, the results regarding the potential effect of number and animacy on overexplicit production are similar to those obtained for the intermediate learners. Although the trend is confirmed (singular and animate referents are more problematic than plural and inanimate) there is also a significant amount of inanimate and plural overexplicit subjects. Consider the example below, extracted from the production of an advanced English-speaking learner:

153) Guevara$_i$ se fue con un amigo$_j$ se llama Alberto Granado para viajar por América del sur. **Ellos$_{ij}$** fueron a una leprosería en Perú. Cuando estuvieron allí, **los dos$_{ij}$** vieron las condiciones sociales de los pobres (ENG38_57_7_2_jd)

"Guevara$_i$ went with a friend$_j$ who is called Alberto Granado to travel through South America. **They$_{ij}$** went to a leper colony in Peru. When (they) were there, **they$_{ij}$** both show the social conditions of the poor people"

According to the criteria that have been established in this study (section 5.5.3), both plural referential choices in example (153) are overexplicit, since in same-reference with only one potential antecedent a null subject would be the pragmatic choice. We will return to this matter after examining the upper-advanced groups in order to conclusively determine the effect of animacy and number for all proficiency levels.

CLAUSE TYPE

We will now focus on the clause types produced by the advanced learners in order to examine how the frequency of overexplicit production may vary in main, coordinate and subordinate clauses. Recall here that all anaphoric subjects were tagged for belonging to one of the aforementioned clause types. In order to compare the frequencies of overexplicit subjects in each type for the English2 and Greek2 groups, two independent 3x2 chi-square tests were performed (one for each group), following the methodological approach of the intermediate learners' section. Starting with the English2 group, in Table 30 we see the results of the statistical analysis (the original UAM CorpusTool raw frequencies can be seen in Figure 90, Figure 91 and Figure 92 in the Appendix):

clause type \ felicity		**Pragmatic subjects**	**Overexplicit subjects**	sum
Main	observed count	108 **(62.06%)**	66 **(37.94%)**	174
	expected count	*123.83*	*50.17*	
	χ2 value	(2.02)	(5)	
Coordinate	observed count	62 **(96.87%)**	2 **(3.12%)**	64
	expected count	*45.55*	*18.45*	
	χ2 value	(5.94)	(14.67)	
Subordinate	observed count	25 **(69.44%)**	11 **(30.56%)**	36
	expected count	*25.62*	*10.38*	
	χ2 value	(0.02)	(0.04)	
sum		195	79	274
χ^2 = 27.685, df = 2, χ^2/df = 13.84 , P(χ^2 > 27.685) = 0.0000				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 30. English2 group: overexplicitness by clause type

Overall, results indicate that there are highly significant differences between the three clause types ($\chi2=27.68$, $p<.0001$). After performing three post-hoc pairwise comparisons (alpha level adjusted at 1.6%) we found that the English2 group employed significantly fewer overexplicit subjects in coordinate than in the other two clause types[126] (Fisher's exact test: $p<.001$ for both coordinate vs main and coordinate vs subordinate comparisons). No differences were found between main and subordinate clauses ($\chi2=0.42$, $p=0.5169$). Crucially, the results for the advanced English2 group are similar to those of the intermediate English1 group (see Table 24 and Table 25). Recall here that null subjects in same-subject coordinate structures are a locus of potential positive cross-linguistic influence, since subject ellipsis in such patterns is allowed in both the L1 (English) and L2 (Spanish) of the learners. Very importantly, the other two clause types do not allow null subjects in English. Consequently, evidence seems to indicate some cross-linguistic influence in the production of the English2 group, in line with the results of the English1 group (the reader is referred to the analysis of this group in section 6.3.1.1 for more details).

Subsequently, the same procedure was followed for the Greek2 group. We start by comparing the frequencies of overexplicit subjects in the three clause types. For that purpose, a 3x2 chi-square test was performed. Results are presented in Table 31 (the original UAM CorpusTool raw frequencies can be seen in Figure 90, Figure 91 and Figure 92 in the Appendix):

[126] Both overexplicit subjects produced by this group concern same-subject coordination. Thus, no further examination per coordination type (same-subject vs different-subject) is required in this case.

clause type \ felicity		Pragmatic subjects	Overexplicit subjects	sum
Main	observed count	125 **(83.33%)**	25 **(16.67%)**	150
	expected count	*131.56*	*18.44*	
	χ2 value	(0.33)	(2.33)	
Coordinate	observed count	52 **(96.29%)**	2 **(3.71%)**	54
	expected count	*48.24*	*6.76*	
	χ2 value	(0.47)	(3.35)	
Subordinate	observed count	35 **(92.10%)**	3 **(7.90%)**	38
	expected count	*34.2*	*4.8*	
	χ2 value	(0.09)	(0.67)	
sum		212	30	242
χ^2 = 7.249, df = 2, χ^2/df = 3.62 , P(χ^2 > 7.249) = 0.0267				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 31. Greek2 group: overexplicitness by clause type

Results demonstrate that likewise the English2 group, in the Greek2 group there is a significant overall correlation between clause type and overexplicitness (χ2=7.25, p=.0267). Due to the low number (less than 5 cases) of overexplicit subjects in some cells, three post-hoc pairwise comparisons with Fisher's exact tests were performed (alpha level adjusted at 1.6%). The results regarding coordinate clauses indicate a similar trend to the one observed for the English1 group, i.e. the production of overexplicit subjects drops. However, the difference between main and coordinate clauses only approximates but does not reach significance (p=.0175). Additionally, neither the difference between coordinate and subordinate clauses nor the difference between main and subordinate clauses is significant (p=.6454 and p=.2092 respectively). The results for the Greek2 group are, in part, similar to those of the Greek1 group (see Table 26). Note, however, that the difference between coordinate and subordinate clauses is inexistent in the advanced group whereas it approximated significance regarding the Greek-speaking group of intermediate proficiency. It has already been pointed out that coordinate structures are triggering the presence of null subjects is both Greek and Spanish (see also 6.2.1). This might explain the lower frequency of overexplicit reference in these contexts. For more details on this matter, the reader is referred to the corresponding analysis of the intermediate Greek1 group (section 6.3.1.1).

PRI

We will now turn to the potential referential interference (PRI) in order to examine how the presence of other potential antecedents may trigger or not the presence of overexplicit

referential subjects. We start by examining the English2 group's production in the four categories of the Active Referents tag. Following the same statistical analysis with the Clause Type category, a 4x2 chi-square test was performed. The results are presented in Table 32 (the original UAM CorpusTool raw frequencies can be seen in Figure 94, Figure 95, Figure 96 and Figure 97 in the Appendix):

felicity / active referents		**Pragmatic subjects**	**Overexplicit subjects**	sum
one_ref	observed count	111 **(63.79%)**	63 **(36.21%)**	174
	expected count	*123.83*	*50.17*	
	$\chi 2$ value	(1.33)	(3.28)	
two_ref	observed count	55 **(83.33%)**	11 **(16.67%)**	66
	expected count	*46.97*	*19.03*	
	$\chi 2$ value	(1.37)	(3.39)	
three_ref	observed count	15 **(78.94%)**	4 **(21.06%)**	19
	expected count	*13.52*	*5.48*	
	$\chi 2$ value	(0.16)	(0.4)	
fourplus	observed count	14 **(93.33%)**	1 **(6.67%)**	15
	expected count	*10.68*	*4.32*	
	$\chi 2$ value	(1.04)	(2.56)	
sum		195	79	274
χ^2 = 13.524, df = 3, χ^2/df = 4.51 , P(χ^2 > 13.524) = 0.0036				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 32. English2 group: overexplicitness and PRI

Overall, the results indicate that there is a significant effect of number of active referents to overexplicitness ($\chi 2$=13.52, p=.0036). Accordingly, due to the low number of overexplicit subjects (<5) in some cells, six post-hoc pairwise comparisons with Fisher's exact tests were performed (alpha level adjusted at 0.8%). In line with the analysis of the English1 group (see Table 27), the results indicate that the English2 group is using significantly more overexplicit subjects with one than with two active referents (p=.0045). However, there is no significant difference between one and three (p=.2154) and one and four referents (p=.0216). Additionally, there are no significant differences between any of the other categories (two vs three: p=.7347, two vs four: p=.4486, three vs four: p=.3547). Recall here that, as we have already noticed, English-speaking learners mostly overproduce redundant pronominals in total absence of any potential ambiguity (same-reference contexts with only one active referent). Therefore, it has been argued that cross-linguistic influence may account for this behaviour. The results of the analysis

for overexplicitness by clause type for the advanced English2 groups partly corroborate this finding. The trend is confirmed whereas the differences are less pronounced. A full account for all groups and proficiency levels will be provided in the last section of the results.

Subsequently, the Greek2 group's production was tested following the exact same procedure as with the English2 group. The frequencies of overexplicit production by number of active referents are shown in Table 33 (the original UAM CorpusTool raw frequencies can be seen in Figure 94, Figure 95, Figure 96 and Figure 97):

active referents \ felicity		**Pragmatic subjects**	**Overexplicit subjects**	sum
one_ref	observed count	157 **(88.70%)**	20 **(11.30%)**	177
	expected count	*155.24*	*21.76*	
	$\chi 2$ value	(0.02)	(0.14)	
two_ref	observed count	41 **(85.41%)**	7 **(14.59%)**	48
	expected count	*43.85*	*6.15*	
	$\chi 2$ value	(0.02)	(0.12)	
three_ref	observed count	11 **(78.57%)**	3 **(21.43%)**	14
	expected count	*12.28*	*1.72*	
	$\chi 2$ value	(0.13)	(0.95)	
fourplus	observed count	3 **(100%)**	0 **(0%)**	3
	expected count	*2.63*	*0.37*	
	$\chi 2$ value	(0.05)	(0.37)	
sum		212	30	242
χ^2 = 1.801, df = 3, χ^2/df = 0.60 , P(χ^2 > 1.801) = 0.6147				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 33. Greek2 group: overexplicitness and PRI

Overall, for the Greek2 group, there are no significant differences between the four conditions ($\chi 2$=1.80, p=.6147). The results indicate that the Greek2 group participants are equally overexplicit for any number of competing active referents in the immediately preceding discourse. This is in line with the results of the intermediate Greek1 group (see Table 28). As it has been suggested, potential ambiguity in terms of numerous antecedent candidates does not seem to trigger overexplicit production. In the case of the advanced Greek-speaking learners, unpragmatic referential choices mostly concern the repeated use of noun phrases instead of the more economical and pragmatic null subject. This might be stylistically explained (use of noun phrases for a more elaborated discourse). Some lingering deficits concerning the production of redundant overt pronouns may be accounted for in terms of processing difficulties, defective input and/or L3 influence.

Crucially, the evidence so far indicates that the combination of several factors may be the only complete explanation for the phenomenon. After obtaining the full picture by examining the last proficiency level (upper-advanced groups) we shall return to this matter in order to fully account for it.

PAS

To conclude, we shall focus on the analysis of the advanced learners on the PAS structures, in line with the procedure followed for the intermediate groups. Recall here that, regarding the English1 and Greek1 groups, the PAS structures could not be tested due to the extremely low number of cases. Concerning the advanced groups, as we can see in Table 34 (the original UAM CorpusTool raw frequencies can be seen in Figure 98 in the Appendix), the PAS structures barely represent 1.5% of the English2 data and 1% of the Greek2 data (overall there are 4 cases in the English2 and 2 cases in the Greek2 data). Overall, the total number of cases is, thus, not sufficient for the application of valid statistical comparisons.

PAS / Group	**PAS not applicable**	**PAS Intersentential**	**PAS Intrasentential**	**Total**
English2	275	3	1	279
	98.57%	**1.07%**	**0.36%**	**100%**
Greek2	244	1	1	246
	99.18%	**0.41%**	**0.41%**	**100%**

Table 34. Advanced learners: PAS structures

Recall here that, as it has been already argued, PAS structures are rather infrequent in the texts examined in this dataset. It might be suggested that this is a side-effect of the specific discourse genre considered here. Notice, however, that the texts under study should be expected to be, a priori, very favourable to the presence of PAS constructions. Narrative and expository essays with several 3rd person referents which are highly interacting with each other seem ideal for testing the PAS hypothesis in real discourse. It is hard to imagine any other discourse genre that would favour more the presence of such structures. Thus, the reason for the scarcity of PAS cases is not likely to be related to the nature of the discourse genre under study. As it has been already suggested, discourse anaphoric structures are generally much more complex than the PAS conditions. It is not surprising,

thus, that we find it hard to obtain a significant amount of cases in order to examine these particular structures in real discourse.

SUMMARY (advanced learners)

In sum, regarding overexplicit production in the anaphoric discourse of the advanced learners, there are some important findings that should be underlined before proceeding with the rest of the analysis. First, both groups of advanced learners (English2 and Greek2) present some deficits concerning the production of overexplicit anaphoric subject expressions in the same discourse patterns that native speakers prefer less explicit forms. This finding, also in line with the results of the intermediate groups, further supports Hypothesis II (a). It should be noted here that English-speaking learners of advanced proficiency are again significantly more overexplicit than their Greek counterparts, who approximate (but not reach) native-like behaviour regarding the phenomenon under study. This finding, in line with the results of the intermediate groups, further supports Hypothesis V regarding the facilitating role of the L1. Secondly, in line with the results of the intermediate groups, advanced learners are less redundant regarding certain grammatical patterns that, in their respective native languages, less explicit forms are also preferred. Both advanced groups approximate native-like performance in the production of null subjects in coordinate clauses. As it has been already suggested, cross-linguistic influence might account for this behaviour, since null subjects are also allowed in coordinate clauses (but crucially not in any of the other clause types examined) in English L1. This provides further support for Hypothesis V. Thirdly, potential referential interference does not promote overexplicitness in advanced learners discourse. Recall here that this was also true for the intermediate learners. The English-speaking group mostly produces redundant pronominals in absence of any other potential competing referents, whereas the Greek-speaking group is equally overexplicit for any number of potential antecedents. It has been suggested that transfer might account for this phenomenon regarding the English group, since overt pronouns are obligatory in English, irrespective of the presence of other competing referents. However, this explanation is not valid for Greek learners whose L1 allows null subjects like Spanish. Although the results of both intermediate and advanced learner groups provide some strong evidence that the L1 is a facilitating factor, a very important question also arises: why do Greek-speaking learners produce redundant subject forms? Overall, the results reported in this section further support Hypothesis VI regarding a multifactorial account of target-deviant performance with anaphoric subjects in Spanish L2. We shall come back to this issue

after examining the last proficiency level (upper-advanced learners). Finally, regarding PAS structures, once again it was not possible to test the pertinent hypotheses due to almost inexistent number of cases in the discourse of both advanced learner groups. In the next section, the present analysis of overexplicitness will be complemented and concluded by examining the upper-advanced learner groups.

6.3.1.3 Overexplicitness in the upper-advanced learners

Focusing on overexplicitness in the upper-advanced learners, the frequencies for each group regarding the production of overexplicit subjects are graphically represented in Figure 33 and reveal important differences between the English3 and the other two groups (the original UAM CorpusTool raw frequencies can be seen in Figure 85 in the Appendix). The English3 learner group produces significantly more overexplicit subjects than the Greek3 group ($\chi2=17.28$, $p<.0001$) and the natives ($\chi2=26.69$, $p<.0001$). Crucially, there is no significant difference between the Greek3 group and the natives ($\chi2=0.01$, $p=0.9203$). Accordingly, the groups are classified according to redundant production in this way: English3 (20.83%) > Greek3 (7.32%), English3 (20.83%) >Natives (7.12%) and Greek3 (7.32%) ≈ Natives (7.12%).

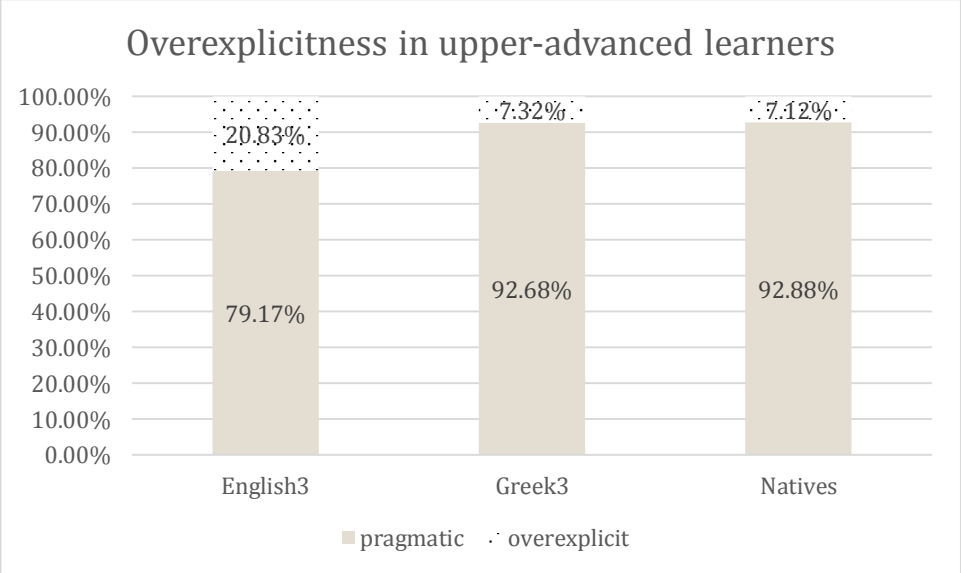


Figure 33. Upper-advanced learners: overexplicitness

Regarding the upper-advanced English3 group, the results are similar to the corresponding intermediate and advanced English groups (see Figure 25 and Figure 29). English-speaking learners produce a significant amount of overexplicit subject forms (20.83%) even at the upper-advanced level of proficiency. Although to a lesser degree

than the corresponding less proficient English1 and English2 groups, the referential choices of the upper-advanced English3 group are significantly more overexplicit than those of the native speakers. Consider the example below, extracted from the production of this group:

154) La protagonista$_i$ se llama Verónica Franco, quien es una joven$_i$ muy hermosa de Venecia. **Ella$_i$** es muy independiente, inteligente, curiosa y un poco marimacho. Al principio de la película, **ella$_i$** se enamora de un hombre (ENG42_21_8_3_LBK)

"The protagonist$_i$ is called Veronica France, who is a young very beautiful girl$_i$ from Venice. **She$_i$** is very independent, intelligent, curious and a little bit tomboy. At the beginning of the film **she$_i$** falls in love with a man"

In example (154) we observe the same kind of constant overexplicit referential choices that have been previously detected in the production of both the intermediate English1 and the advanced English2 groups. In the above discourse passage, although there is only one active referent ("the protagonist"), the writer chooses an overexplicit overt pronominal for reference maintenance. Note that a null subject is pragmatically more appropriate in these contexts. It has been already suggested that cross-linguistic influence may account for this persisting linguistic behaviour of the English-speaking learners of Spanish. Recall that null subjects are ungrammatical in English in the vast majority of discourse patterns under study. The influence of L1 seems to significantly constrain the anaphoric choices of English learners even at very high proficiency levels. As a result, although being grammatically correct, their discourse production is not native-like due to the pragmatically inappropriate anaphoric subjects. For more details, the reader is referred to the analysis of the English1 and English2 groups in sections 6.3.1.1 and 6.3.1.2 respectively.

On the other hand, the upper-advanced Greek3 participants produce less redundant forms than their English3 counterparts and, crucially, demonstrate native-like frequencies of overexplicitness. Both the Greek3 and the native control group are only sporadically overexplicit (around 7% of the total number of anaphors, for both groups). Taking into account that, so far, all learner groups have exhibited an improvement towards native performance and that the advanced Greek2 group was already only marginally more overexplicit than the native group, this is a rather expected finding at this point of the analysis. Yet, it is crucial for the research questions of this study. On one side, this finding suggest that some properties regarding the anaphoric distribution of subject forms can be fully acquired. Additionally, in conjunction with the results of the intermediate and advanced groups, this finding provides solid evidence that the L1 is a facilitating factor

regarding the acquisition of 3rd person anaphoric subjects in Spanish L2. All Greek-speaking learner groups perform significantly better than the English-speaking same-proficiency groups. In addition, the upper-advanced Greek3 group, in contrast to the English3 group, reaches native-like performance regarding the production of pragmatically appropriate subject forms. This matter shall be broadly considered in the summary of this section and in the general discussion.

With the purpose of accounting for the above findings in a more detailed manner we will now examine the type of overexplicit subjects produced by each group. In line with the procedure followed for the intermediate and advanced learners, overexplicitness was classified into three types, according to the actual referential choice and the expected form of the anaphoric expression (as described in section 5.5.3). Figure 34 shows the distribution of overexplicit subjects by type for the three groups two upper-advanced learner groups (English3 and Greek3) and the native speakers (the original UAM CorpusTool raw frequencies can be seen in Figure 85 in the Appendix):

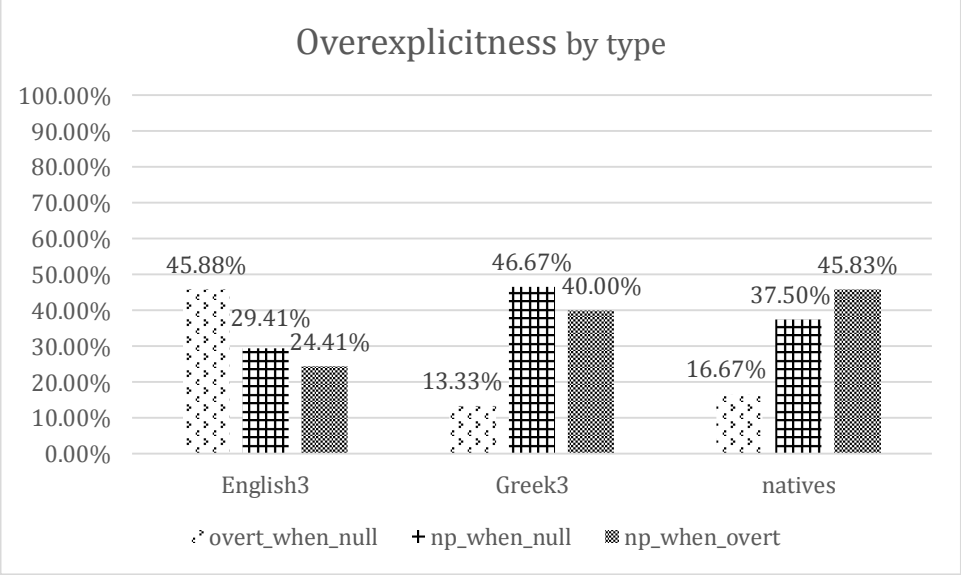


Figure 34. Upper-advanced learners: overexplicitness by type

Regarding the English3 group, we observe that the pattern of overexplicitness by type is similar to the one that was observed for the English1 and English2 groups (see Figure 26 and Figure 30). The upper-advanced English-speaking learners produce an important amount (45.88%) of unpragmatic pronominal subjects (‘overt when null’ category), in cases where a null subject would be sufficiently informative. Two Fisher’s exact tests were performed (due to the low number of redundant overt pronouns in the other two groups) and revealed that English3 group differs significantly from the Greek3 group

(p=.0222) and the natives (p=.0167) to this respect. Crucially, there are no significant differences between the upper-advanced English group and the other two groups for any other type of overexplicitness. More specifically, for 'np_when_null', the English3 group does not differ neither from the Greek3 (p=.2324) nor from the natives (p=.4631). Similarly, regarding the 'np_when_overt' type of overexplicitness, the English3 groups does not differ from the Greek3 group (p=.3432) nor the natives (p=.0734).

On the other side, the pattern of overexplicitness by type for the upper-advanced Greek3 group is quite similar to the one of the native speakers. For both groups, their scarce overexplicit production mostly concerns noun phrases and only some sporadic overt pronominals. Three Fisher's exact tests revealed no significant differences for any type of overexplicitness between the Greek3 and the control group (for 'overt_when_null': p=1, for 'np_when_null': p=.7397 and for 'np_when_overt': p=.7526). Recall here that the production of noun phrases in contexts where less explicit forms would be sufficiently informative, though technically overexplicit, may entail some stylistic purposes. The same is not true regarding overexplicit pronominals. This matter has been already discussed in the analysis of the intermediate groups (see section 6.3.1.1 for more details). The distributions in Figure 34 further support the hypothesis that the production of the English3 group is, to some degree, constrained by the grammatical features of English. On the contrary, the similar distribution regarding type of overexplicitness between the Greek3 and the native group further confirm the potential facilitating effect of the Greek learners' L1.

We shall now focus exclusively on the two upper-advanced learner groups in order to independently examine their production of overexplicit subjects in relation to other discourse factors. In line with the procedure followed for the intermediate and advanced groups, the following factors will be considered: number, animacy, clause type and PRI. Furthermore, potential deficits in PAS contexts will be separately analysed. Each group's production will be first examined on its own and then compared to the production of the other group.

NUMBER

With respect to the grammatical number, Figure 35 demonstrates the frequency of overexplicit choices for singular and plural subjects for the two upper-advanced learner groups (the original UAM CorpusTool raw frequencies can be seen in Figure 86 and Figure 87 in the Appendix):

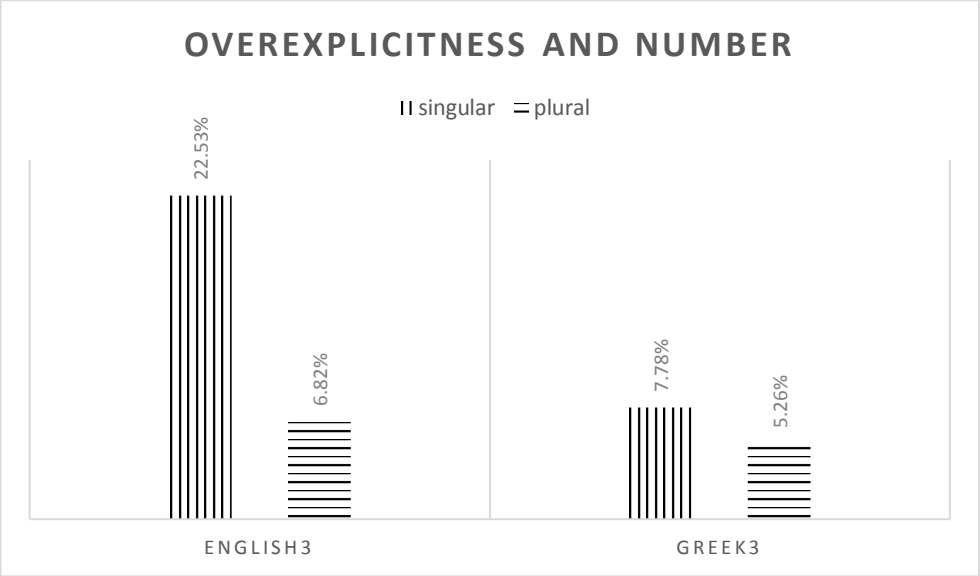


Figure 35. Upper-advanced learners: overexplicitness and grammatical number

In line with the results of the intermediate and advanced groups, we observe that for both upper-advanced learner groups, 3rd person singular referents are more problematic than the plural ones. Additionally, in contrast with the corresponding results of the less proficient groups, the difference between the frequency of overexplicit subjects in singular and plural number is significant for the English3 group ($\chi 2=4.96$, $p=.0167$). No significant difference is found regarding the anywise scarce overexplicit production of the Greek3 group (Fisher's exact test: $p=.742$). We can thus conclude, in line with Lozano (2009b), that the upper-advanced English-speaking learners are significantly more overexplicit with singular than with plural 3rd person anaphoric subjects. Their Greek counterparts, on the other side, produce sporadically some overexplicit subjects in both grammatical numbers. Recall here that the exact same trend has been observed for all learner groups (i.e., more overexplicit anaphors in singular than in plural number). However, only for the upper-advanced English3 group did the differences reach statistical significance.

ANIMACY

Next we shall turn our attention to the potential effect of animacy in the production of overexplicit subjects for the upper-advanced learner groups. In Figure 36 we see that both groups are more overexplicit with animate than with inanimate referents (the original UAM CorpusTool raw frequencies can be seen in Figure 88 and Figure 89 in the Appendix):

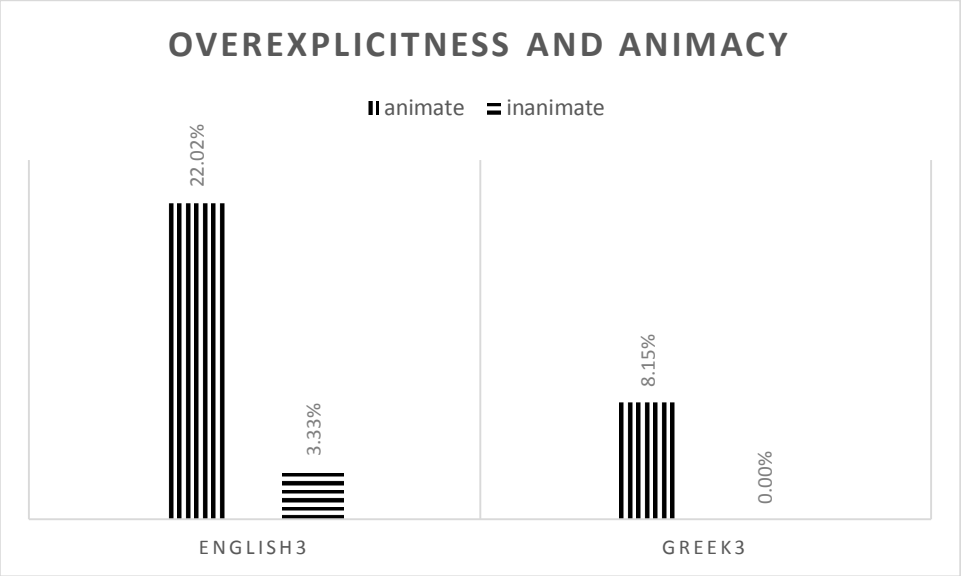


Figure 36. Upper-advanced learners: overexplicitness and animacy

The difference, regarding the English3 group, is statistically significant (Fisher's exact test: p=.0167). On the other hand, the scarce overexplicit production of the Greek3 group does not significantly depend on the animacy of the referent (Fisher's exact test, p=.2386). Overall, results indicate that overexplicitness in the English-speaking upper-advanced learners concerns almost exclusively animate referents. These results are in line with the findings of Lozano (2009b). The scarce overexplicit production of the Greek3 group, on the other hand, is not found to depend on the animacy of the referent. Overall, the results regarding animacy confirm the general trend (though not reaching significance) that has been observed for all learner groups: more overexplicit subject forms with animate than with inanimate referents. This trend seems to be especially relevant for the upper-advanced English3 group, where the difference is particularly pronounced and reaches statistical significance.

CLAUSE TYPE

We shall now focus on the clause types produced by the upper-advanced learners in order to examine how the frequency of overexplicit production may vary depending on whether the subject form belongs to a main, coordinate or subordinate clause. Recall here that all anaphoric subjects were tagged for belonging to one of the aforementioned clause types. In order to compare the frequencies of overexplicit subjects in each type for the English3 and Greek3 groups, two independent 3x2 chi-square tests were performed (one for each group), following the same methodological approach used for the intermediate and advanced learners. Starting with the English3 group, in Table 35 we see the results of the

statistical analysis (the original UAM CorpusTool raw frequencies can be seen in Figure 90, Figure 91 and Figure 92 in the Appendix):

clause type \ felicity		**Pragmatic subjects**	**Overexplicit subjects**	sum
Main	observed count	146 **(74.11%)**	51 **(25.89%)**	197
	expected count	*155.96*	*41.04*	
	$\chi 2$ value	(0.64)	(2.42)	
Coordinate	observed count	90 **(84.11%)**	17 **(15.89%)**	107
	expected count	*84.71*	*22.29*	
	$\chi 2$ value	(0.33)	(1.26)	
Subordinate	observed count	87 **(83.65%)**	17 **(16.35%)**	104
	expected count	*82.33*	*21.67*	
	$\chi 2$ value	(0.26)	(1.01)	
sum		323	85	408
$\chi^2 = 5.909$, df = 2, χ^2/df = 2.95 , P($\chi^2 > 5.909$) = 0.0521				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 35. English3 group: overexplicitness by clause type

Overall, the results indicate that the differences between the three clause types regarding overexplicit production are on the margins of significance ($\chi 2$=5.91, p=.0521). After performing three post-hoc pairwise comparisons (alpha level adjusted at 1.6%) we found that, for the English3 group, the differences between the three types of clauses are not significant (main vs coordinate: $\chi 2$=3.44, p=.0636, main vs subordinate: $\chi 2$=3.02, p=.0822 and coordinate vs subordinate: $\chi 2$=0.01, p=1). Recall, however, that same-subject coordination is of particular interest in order to test for potential cross-linguistic influence (see 6.3.1.1 for more details on the two types of coordinate clauses). In order to achieve a more precise account, switch-reference coordinate clauses were excluded from the analysis and an additional comparison was performed. The results are shown in Table 36 (the original UAM CorpusTool raw frequencies can be seen in Figure 93 in the Appendix):

clause type \ felicity		Pragmatic subjects	Overexplicit subjects	sum
Main	observed count	146 **(74.11%)**	51 **(25.89%)**	197
	expected count	*158.12*	*38.88*	
	χ2 value	0.93	3.78	
Coordinate (only same-subject)	observed count	72 **(92.30%)**	6 **(7.69%)**	78
	expected count	*63.41*	*15.59*	
	χ2 value	1.16	4.73	
Subordinate	observed count	87 **(83.65%)**	17 **(16.35%)**	104
	expected count	*83.47*	*20.53*	
	χ2 value	0.15	0.61	
sum		305	74	379
χ^2 = 11.360, df = 2, χ^2/df = 5.68 , P(χ^2 > 11.360) = 0.0034				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 36. English3 group: overexplicitness in same-subject coordinate clauses

Crucially, the difference between main and coordinate (same-subject) clauses was found to be significant (χ2=10.18, p=.0014). On the other hand, no significant differences were found between main and subordinate clauses (χ2=3.02, p=.0822) nor between coordinate and subordinate clauses (χ2=2.29, p=.1302). At first sight, the results of the upper-advanced English3 group regarding a potential clause type effect are slightly different from those of the intermediate and advanced English groups. Recall here that null subjects in coordinate structures were previously suggested to be a locus of potential cross-linguistic influence, since subject ellipsis in such patterns is allowed in both the L1 (English) and L2 (Spanish) of the learners. Although there is a clear trend concerning the decreased frequency of overexplicitness in coordination for all groups, the evidence for the upper-advanced group indicate that cross-linguistic influence may not be the only relevant factor in the production of the English-speaking learners. At least, not to the same extent for all proficiency levels, since the above findings concerning the upper-advanced group indicate a diminished effect for this factor. We shall come back to this matter in the general discussion.

Subsequently, the same procedure was followed for the Greek3 group. We start by comparing the frequencies of overexplicit subjects in the three clause types. For that purpose, a 3x2 chi-square test was performed. Results are presented in Table 37 (the original UAM CorpusTool raw frequencies can be seen in Figure 90, Figure 91 and Figure 92 in the Appendix):

clause type \ felicity		Pragmatic subjects	Overexplicit subjects	sum
Main	observed count	98 **(89.09%)**	12 **(10.91%)**	110
	expected count	*101.95*	*8.05*	
	$\chi 2$ value	(0.15)	(1.94)	
Coordinate	observed count	54 **(98.18%)**	1 **(1.82%)**	55
	expected count	*50.98*	*4.02*	
	$\chi 2$ value	(0.18)	(2.27)	
Subordinate	observed count	38 **(95%)**	2 **(5%)**	40
	expected count	*37.07*	*2.93*	
	$\chi 2$ value	(0.02)	(0.29)	
sum		190	15	205
χ^2 = 4.862, df = 2, χ^2/df = 2.43 , P(χ^2 > 4.862) = 0.0880				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 37. Greek3 group: overexplicitness by clause type

Overall, the results of the upper-advanced Greek3 participants are similar to those of the English3 group. Although the Greek-speaking learners of very high proficiency are more overexplicit in main than in coordinate or subordinate clauses, the differences are overall not significant ($\chi 2$=4.86, p=.0880). Note, however, that the same trend that was observed for all learner groups concerning the reduced overexplicit production in coordinate structures is also found here. In line with the results of the upper-advanced English3 learner group, this might suggest that the effect of cross-linguistic influence might decrease as proficiency level grows. Consequently, the source of overexplicit production for high proficiency groups should be sought elsewhere. A full developmental account for all groups will be given in the last section of the results in order to provide some answers regarding this matter.

PRI

We shall now turn to the PRI factor in order to examine how the frequency of overexplicit referential subjects may depend on the presence of other potential antecedents. The same procedure that was previously used for the intermediate and advanced groups will be followed here. We start by examining the English3 group's production in the four categories of the Active Referents tag. Following the same statistical analysis with the Clause Type category, a 4x2 chi-square test was performed. The results are presented in Table 38 (the original UAM CorpusTool raw frequencies can be seen in Figure 94, Figure 95, Figure 96 and Figure 97):

active referents \ felicity		Pragmatic subjects	Overexplicit subjects	sum
one_ref	observed count	166 **(80.97%)**	39 **(19.03%)**	205
	expected count	*162.29*	*42.71*	
	$\chi 2$ value	(0.08)	(0.32)	
two_ref	observed count	96 **(71.11%)**	39 **(28.89%)**	135
	expected count	*106.88*	*28.12*	
	$\chi 2$ value	(1.11)	(4.21)	
three_ref	observed count	34 **(82.92%)**	7 **(17.08%)**	41
	expected count	*31.67*	*8.33*	
	$\chi 2$ value	(0.17)	(0.65)	
fourplus	observed count	27 **(100%)**	0 **(0%)**	27
	expected count	*22.17*	*5.83*	
	$\chi 2$ value	(1.05)	(4)	
sum		322	86	408
χ^2 = 11.602, df = 3, χ^2/df = 3.87 , P(χ^2 > 11.602) = 0.0089				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 38. English3 group: overexplicitness and PRI

Overall, the results indicate that there is a significant effect of number of active referents to overexplicitness ($\chi 2$=11.60, p=.0089). Accordingly, in order to precisely detect and fully account for the differences, six post-hoc pairwise comparisons were performed (alpha level adjusted at 0.8%). Regarding the 'one_ref' condition, no differences were found with the other three conditions (for 'two_ref': $\chi 2$=3.94, p=.0471, for 'three_ref': $\chi 2$=0.01, p=.9203 and for 'fourplus': p=.0107). Similarly, no differences were found between the 'two_ref' and the 'three_ref' conditions ($\chi 2$=1.7, p=.1922) nor between the 'three_ref' and the 'four_plus' condition (p=.0366). A significant difference was found only between the 'two_ref' and the 'four_plus' condition (p=.0026). Recall here that the English-speaking intermediate and advanced learner groups were found to mostly overproduce redundant pronominals in total absence of any potential ambiguity (same-reference contexts with only one active referent). On the other side, the upper-advanced group's overexplicit production seems to be, at least in part, of different nature. Therefore, as it has already been argued, cross-linguistic influence may not fully account for this behaviour. The results of the present analysis are also in line with the findings of the previous section (overexplicitness by clause type).

Subsequently, the Greek3 group's production was examined following the exact same procedure with the English3 group. The frequencies of overexplicit production by number

of active referents are demonstrated in Table 39 (the original UAM CorpusTool raw frequencies can be seen in Figure 94, Figure 95, Figure 96 and Figure 97):

active referents \ felicity		**Pragmatic subjects**	**Overexplicit subjects**	sum
one_ref	observed count	104 **(92.85%)**	8 **(7.15%)**	112
	expected count	*103.8*	*8.2*	
	χ2 value	(0)	(0)	
two_ref	observed count	42 **(89.36%)**	5 **(10.64%)**	47
	expected count	*43.56*	*3.44*	
	χ2 value	(0.06)	(0.71)	
three_ref	observed count	24 **(92.30%)**	2 **(7.70%)**	26
	expected count	*24.1*	*1.9*	
	χ2 value	(0)	(0.01)	
fourplus	observed count	20 **(100%)**	0 **(0%)**	20
	expected count	*18.54*	*1.46*	
	χ2 value	(0.12)	(1.46)	
sum		190	15	205
χ^2 = 2.354, df = 3, χ^2/df = 0.78 , P(χ^2 > 2.354) = 0.5023				
expected values are displayed in *italics* individual χ^2values are displayed in (parentheses)				

Table 39. Greek3 group: overexplicitness and PRI

Overall, for the Greek3 group, there are no significant differences between the four conditions (χ2=2.35, p=.5023). This result is in line with the results of the Greek1 and Greek2 groups (see Table 28 and Table 33) and is rather expected given the overall low number of overexplicit anaphoric subjects in the production of the upper-advanced Greek learner group. It further suggests that overexplicitness is not triggered by the presence of more potential antecedents.

PAS

Finally, the analysis of the upper-advanced groups regarding overexplicit production will be concluded by considering the PAS hypothesis. In line with the procedure followed for the intermediate and advanced learners, PAS structures shall be separately analyzed. Recall here that for the lower proficiency groups it was not possible to examine the PAS hypothesis due to the very low number of cases in the dataset. In Table 40 we observe the results of the upper-advanced groups (the original UAM CorpusTool raw frequencies can be seen in Figure 98 in the Appendix):

PAS / Group	PAS not applicable	PAS Intersentential	PAS Intrasentential	Total
English3	413	5	5	423
	97.64%	**1.18%**	**1.18%**	**100%**
Greek3	206	5	2	213
	96.71%	**2.35%**	**0.94%**	**100%**

Table 40. Upper-advanced learners: PAS structures

The results of the upper-advanced learner groups further confirm the scarcity of PAS structures in real discourse that was also observed in the production of the intermediate and advanced groups. Only 10 cases, accounting for the 2.36% of their data, were found in the texts of the English3 group. The Greeek3 group produced a total of 7 cases, a number that accounts for the 3.29% of their data. Therefore, it may be concluded that the immense majority of the anaphoric discourse produced by all learner groups of this study (intermediate, advanced and upper-advanced) is not susceptible to be tested for PAS. This finding poses some questions regarding the predictive power of PAS in the anaphoric distribution of 3rd person subjects. The reader is referred to the discussion in the corresponding sections of the intermediate and advanced learners for more details regarding the complexity of discourse anaphoric relations as opposed to the artificial simplification of the PAS constructions.

SUMMARY (upper-advanced learners)

In sum, the results regarding the overexplicit anaphoric production of the upper-advanced learner groups have confirmed an important finding that was already observed in the analysis of the less proficient learners. Overall, Greek learners of Spanish are less redundant than the English learners, for all proficiency levels. Crucially, the upper-advanced Greek3 group does not differ from the native Spanish control group. In SLA terms, there are two direct implications regarding these findings. Firstly, the similarity in the distribution of 3rd person anaphoric subjects between Spanish and Greek seems to be a facilitating factor for the Greek-speaking learners, who may positively transfer the properties of their L1 into their L2 discourse production. Greek learners seem to take advantage of these properties and perform significantly better than their English counterparts at all proficiency levels. This further confirms Hypothesis V regarding the facilitating role of the L1. Conversely (but in a parallel direction), the English learners seem to be negatively constrained by the properties of their corresponding L1 that is, at

least in part, responsible for the significant amount of redundant subject forms in their production. Secondly, anaphoric subjects may not be a locus of persistent deficits for some high proficient learner groups. This finding runs against Hypothesis II (b) which predicted that deficits would persist even at the upper-advanced levels of proficiency. The Greek3 upper-advanced group is completely native-like regarding overexplicitness in the anaphoric patterns that were examined. The results of this section will be complemented with the analysis of the underexplicit production of the learner groups, performed in the next sections.

6.3.1.4 Underexplicitness in the intermediate learners

In order to examine underexplicitness and in line with the procedure followed in the previous sections, overexplicit subjects were excluded from the dataset. Recall that overexplicit discourse patterns are, by definition, the opposite of underexplicitness due to the fact that a redundant form cannot be ambiguous. After extracting overexplicit subjects, the underexplicit referential choices for the intermediate proficiency groups (English1 and Greek1) were analysed and contrasted to the corresponding choices of the native control group. The results for all three groups, in Figure 37, reveal some differences between the native group and the two learner groups (the original UAM CorpusTool raw frequencies can be seen in Figure 99 in the Appendix).

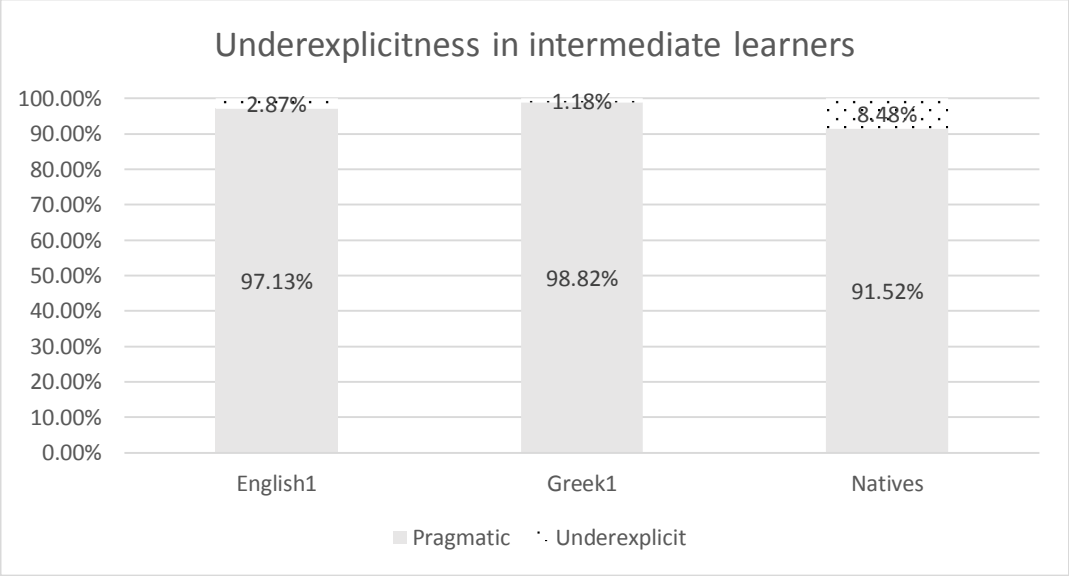


Figure 37. Intermediate learners: underexplicitness

Three Fisher's exact tests were performed (due to the low number of underexplicit subjects for the learner groups) and revealed that native speakers produce significantly

more underexplicit subjects than the English1 (p=.0227) and the Greek1 (p=.0012) groups. No differences were found between the two learner groups in their scarce production of underexplicit subjects (p=.4485). In order to fully account for this phenomenon in the native group, we focus on the four types of underexplicit subjects as described in section 5.5.3. The proportions for each type are graphically represented in Figure 38 (the original UAM CorpusTool raw frequencies can be seen in Figure 99 in the Appendix):

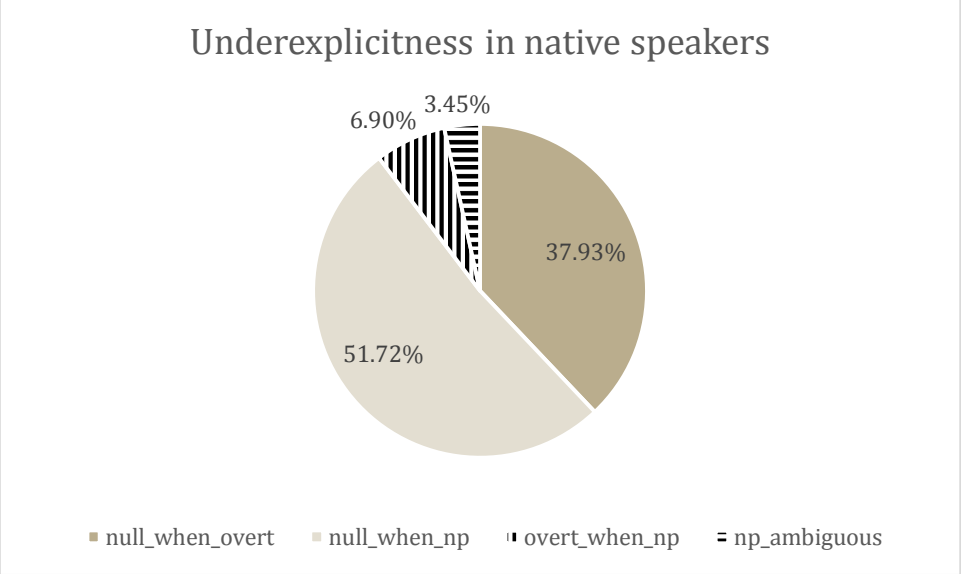


Figure 38. Natives: underexplicitness by type

As we can see in Figure 38, the vast majority of underexplicit subjects in the native group belong to the 'null_when_np' and 'null_when_overt' categories. Regarding the former type, it accounts for nulls subjects produced in discourse patterns where two same-gender referents are active, creating a potentially ambiguous anaphoric context where a noun phrase is expected in order to resolve the anaphoric relation. However, we notice that an alternative way to correctly establish the coreference relation is by making use of the information which the writer assumes to be shared with the reader (previous discourse and world knowledge). Consider the following example:

-unpragmatic/underexplicit/null_when_np

155) Para llegar hasta ella$_i$, Ø$_j$ deberá cruzar un gran muro de piedra que está vigilado por un guardián$_k$ que no permite el paso a nadie. Cuando al fin **Ø$_j$** consigue superar la barrera, Ø$_j$ descubre que al otro lado del muro se encuentra un mundo mágico (ESP21_3_CPV)

"To reach her$_i$, (he)$_j$ must cross a big stone wall which is being watched over by a guard$_k$ who does not allow anybody to pass. When (**he**)$_j$ finally manages to get over the barrier, (he)$_j$ discovers that at the other side of the wall there is a magic world"

In example (155), both "he" and "the guard" (and even "her) could potentially corefer with the null subject of the clause under question ("finally manages to get over the barrier"). Taking into consideration world knowledge though, it is obvious that the one guarding the wall is less likely to try and manage to get over it, while at the same time it is known from previous discourse that "he" must cross the barrier in order to get to his beloved. In this example we can observe how world knowledge interacts with previous discourse and the technical ambiguity is resolved. Additionally, another factor that might possibly promote the activation of the first referent ("he") versus the second ("the guard") is protagonisthood. The referent that "manages to get over the barrier" is the protagonist of the text, thus, less explicit referential choices might be sufficient in order to unambiguously refer to him.

The second most frequent type of underexplicit referential choice ('null_when_overt') concerns contexts where two different-gender referents are active. Technically, a null subject was tagged as ambiguous in such contexts (see section 5.5.3). Consider the following case:

-unpragmatic/underexplicit/null_when_overt

156) Al final ella$_i$ y el príncipe$_j$ van a un baile temático donde **Ø$_i$** se encuentra con su$_i$ salvador (ESP21_3_CPV)

"Finally, she$_i$ and the prince$_j$ are going to a thematic dance where (**she$_i$**) finds her$_i$ rescuer"

In example (156), it is equally plausible that "she" (Cinderella) or "the prince" find a "rescuer". Crucially, the possessive pronoun after the verb is not marked for gender in Spanish. The null subject is, thus, technically ambiguous. Nevertheless, we already know from the previous discourse that the princess is in danger, and that somebody (a lawyer) is going to rescue her. Although an overt feminine pronoun would be the perfectly clear referential choice, the native speaker chooses an underexplicit null subject, since previous discourse diminishes any potential ambiguity. Again, the fact that the referential form concerns the protagonist of the text (the story is about Cinderella) might also contribute to the selection of less explicit forms without risking ambiguity, due to the privileged activation state of protagonist referent.

A very similar explanation applies to the third most frequent underexplicitness type ('overt_when_np'), where an overt pronoun is used in contexts that require a noun phrase. Consider the following example:

-unpragmatic/underexplicit/overt_when_np

157) Cuando $Ø_i$ se dio la vuelta un hombre$_j$ vestido de negro lo$_i$ golpio dejadolo$_i$ inconsiente en el piso. En unos minutos llego la policia, **el$_i$**, todavia inconsiente, miro de reojo que su mujer estaba muerta (ESP16_3_Bv)

"When (he)$_i$ turned around a man$_j$ dressed in black hit him$_i$ and left him$_i$ unconscious in the apartment. In a few minutes the police arrived, **he$_i$**, still unconscious, looked out of the corner of his eye that his wife was dead"

In example (157) an overt masculine pronoun is used to corefer with one of the two active same-gender referents. This case is similar to the 'null_when_np' type, with the only difference that an overt pronoun (instead of a null subject) is employed here. Although the masculine personal pronoun in a context with two masculine activated referents is potentially ambiguous, shared discourse knowledge resolves this technical ambiguity. According to the information in the previous discourse only one of the characters fell unconscious to the floor, and thus he is the only one that can be "still unconscious". Once again, the protagonist status might act as a confounding factor by granting the referent, a priori, a higher activation status which renders it more susceptible to be recovered with less explicit forms.

In order to examine the degree of the association between shared knowledge (previous discourse and world knowledge) and underexplicitness in the native group, the total number of underexplicit subjects was tested for the presence or absence of shared knowledge constraints (see section 5.5.2.9).

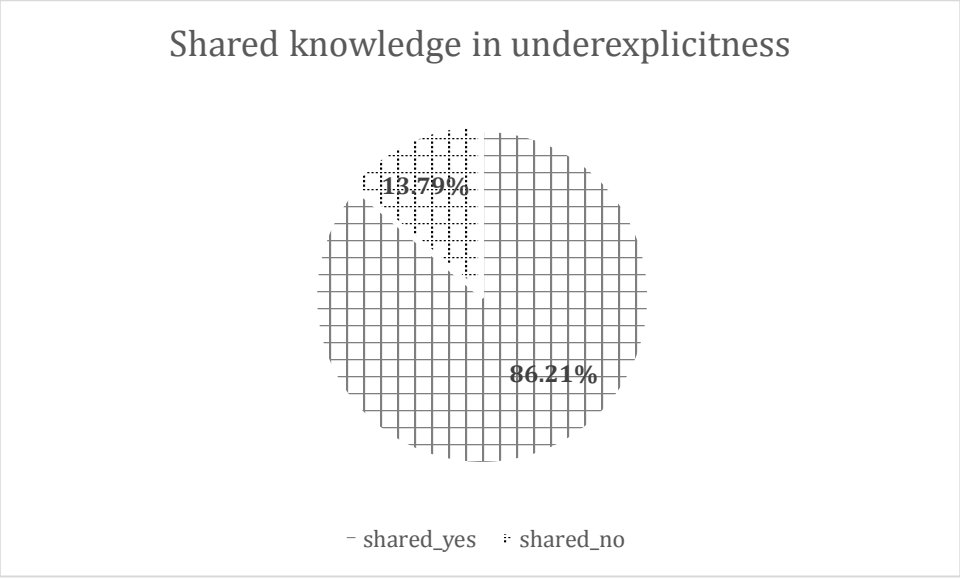


Figure 39. Natives: shared knowledge constraints in underexplicit subjects

The results in Figure 39 (the original UAM CorpusTool raw frequencies can be seen in Figure 101 in the Appendix) show that the vast majority of underexplicit referential choices (86.21%) are constrained by some shared knowledge (previous discourse and/or world knowledge) which impedes ambiguity. Native speakers seem to take into consideration these constraints and produce a considerable proportion of underexplicit subject expressions (see Figure 37). Intermediate learners, to the contrary, prefer to avoid ambiguity at all costs, producing a significant amount of overexplicit discourse (as we saw in section 6.3.1.1) and very few underexplicit subjects. In order to fully account for the importance of shared knowledge, the frequencies of underexplicit and pragmatic subjects that are constrained by shared knowledge were compared with a Fisher's exact test. The results in Table 41 (the original UAM CorpusTool raw frequencies can be seen in Figure 100 and Figure 101 in the Appendix) reveal that significantly more underexplicit than pragmatic subjects are constrained by shared knowledge ($p<.0001$).

Shared Knowledge / Pragmaticality	**Shared_yes**	**Shared_no**	**Total**
Pragmatic subjects	41	272	313
	13.79%	**86.21%**	**100%**
Underexplicit subjects	25	4	29
	86.90%	**13.10%**	**100%**

Table 41. Presence of Shared Knowledge constraint (natives)

It has been further suggested that the protagonisthood factor might have an additional impact on underexplicit referential choices. The results do not confirm this hypothesis, since in Figure 40 we observe that only half of the underexplicit subjects (51.72%) refer to the protagonist of the text (the original UAM CorpusTool raw frequencies can be seen in Figure 101 in the Appendix):

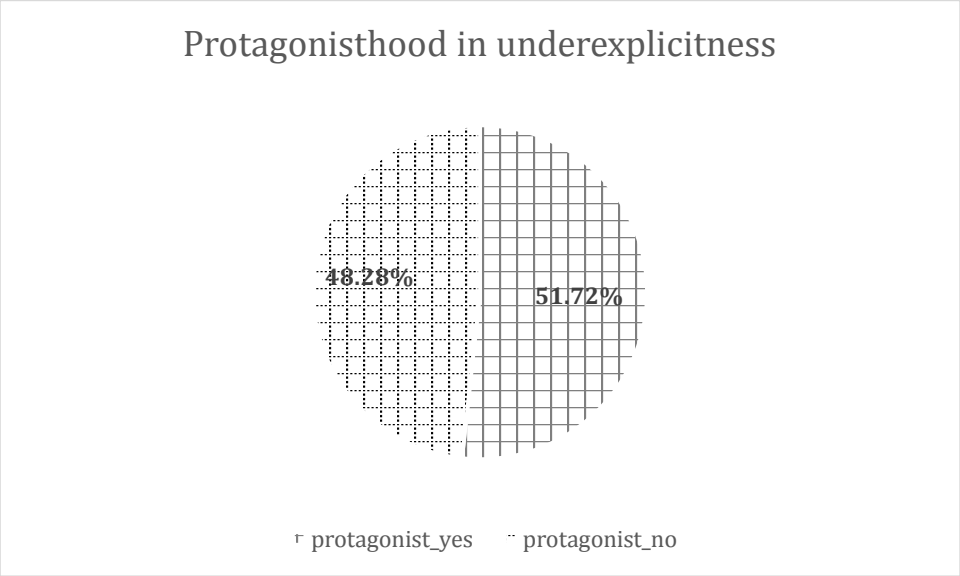


Figure 40. Natives: protagonisthood factor in underexplicit subjects

In sum, regarding the intermediate learners of this study, a very scarce underexplicit production has been detected. To the contrary, as we saw in section 6.3.1.1, both intermediate learner groups produced a significant amount of redundant discourse, to such an extent that no ambiguity should be expected whatsoever. All in all, underexplicitness results are in line with the findings regarding overexplicitness and further support Hypothesis III (a). On the other side, Hypothesis III (b) is not confirmed, insofar as intermediate learners slightly differ from the native speakers regarding underexplicitness, but not to the expected direction. Native speakers, and not learners, produce more underexplicit anaphoric subjects which render their discourse, technically, more ambiguous. In the next section we shall examine the advanced groups in order to explore how proficiency level may have an effect on the observed trends regarding referential choices of the English and Greek learners of Spanish.

6.3.1.5 Underexplicitness in the advanced learners

The overall frequencies of underexplicitness for the advanced learner groups (English2 and Greek2) and the native control group are graphically represented in Figure 41 (the original UAM CorpusTool raw frequencies can be seen in Figure 99 in the Appendix). In line with the results of the intermediate groups, some differences between the advanced learner groups and the native speakers can be observed.

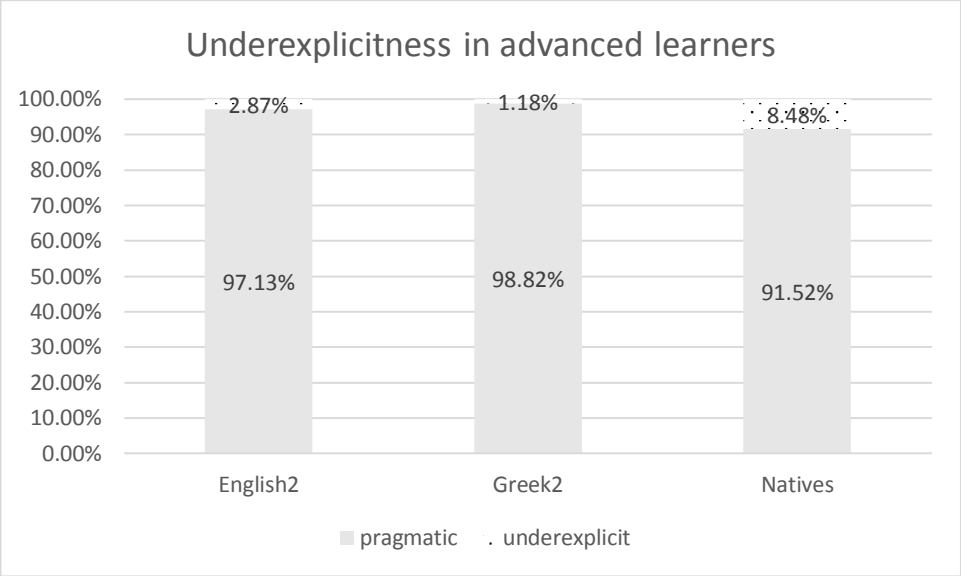


Figure 41. Advanced learners: underexplicitness

Three Fisher's exact tests were performed (due to the low number of underexplicit subjects for the learner groups) and revealed that native speakers produce significantly more underexplicit subjects than the English2 (p=.0053) and the Greek2 (p=.0013) groups. On the other hand, no differences were found between the two learner groups in their scarce production of underexplicit subjects (p=.7433). Recall here, however, that underexplicitness in the native group does not entail unsolvable ambiguity and that it is strongly associated to 'shared information' (previous discourse and/or world knowledge). For the analysis of the native group's underexplicit production the reader is referred to the discussion in the previous section.

Overall, the results for the two advanced learner groups regarding underexplicitness are in line with the findings in the previous section concerning the intermediate groups. Additionally, as we saw in section 6.3.1.2, both advanced learner groups produce a significant amount of redundant discourse, in contrast with the native speakers who take advantage of other discourse features in order to produce optimally elliptical anaphoric patterns. Shared discourse and/or world knowledge favour the use of technically ambiguous referential choices in the native speakers' discourse. The advanced learners, on the other side, do not make use of such sophisticated referential mechanisms. Inversely, they exhibit a more conservative discursive behaviour by making at least sufficiently explicit and, in many cases, overexplicit referential choices. This supports Hypothesis II (a) and does not seem to vary between intermediate and advanced proficiency level for neither the English-speaking nor the Greek-speaking participants.

Hypothesis II (b) is not confirmed, insofar as the advanced learners do not overproduce ambiguous anaphoric subjects. In the next section we shall examine the upper-advanced groups in order to complete the analysis of the learner groups and further explore how proficiency level may have an effect on the observed trends regarding the underexplicit referential choices of the English and Greek learners of Spanish.

6.3.1.6 Underexplicitness in the upper-advanced learners

The results for the English3 and Greek3 groups are graphically represented in Figure 42 (the original UAM CorpusTool raw frequencies can be seen in Figure 99 in the Appendix):

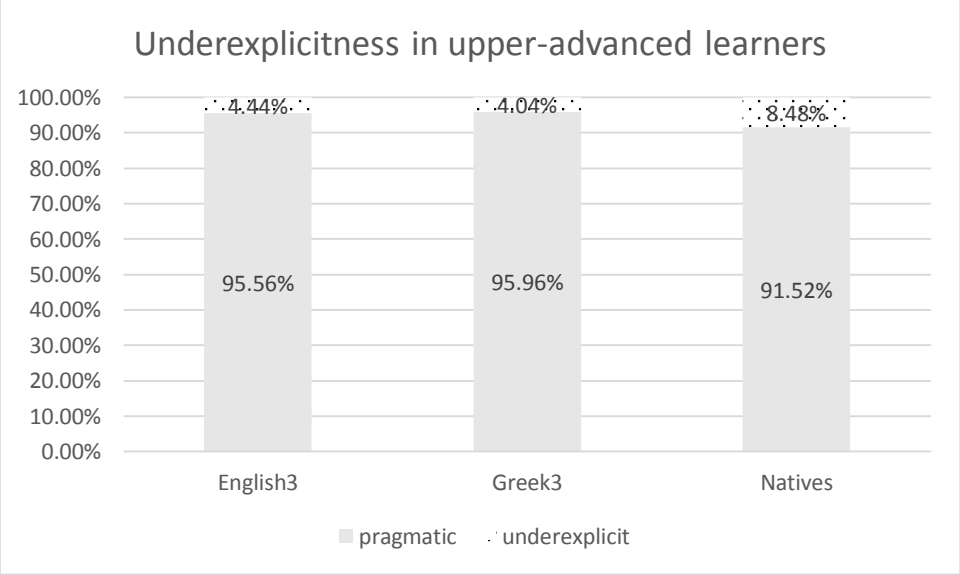


Figure 42. Upper-advanced learners: underexplicitness

In line with the procedure followed in the previous sections, the performance of the English3 and Greek3 learner groups regarding underexplicitness was contrasted to the native speakers. The results revealed that the differences are on the margins of significance (for the English3: $\chi2=3.94$, $p=.0471$ and for the Greek3 $\chi2=3.21$, $p=.0731$). No differences were found between the two learner groups ($\chi2=0.05$, $p=.8265$). The results for the upper-advanced learner groups slightly differ from the findings regarding the production of the intermediate and advanced learners. The more proficient groups seem to approximate the frequencies of underexplicit production of the native control group. It should be reminded here that, although technically ambiguous, the underexplicit anaphoric choices of the natives were found to be strongly correlated to shared information and resolved by means of previous discourse and/or world knowledge. In order to examine if the upper-advanced learners make also use of this sophisticated

referential mechanisms, the underexplicit production of the English3 and Greek3 groups was tested for potential association with the aforementioned factor (see 6.3.1.4 for the same analysis applied to the natives). The results are graphically demonstrated in Figure 43 (the original UAM CorpusTool raw frequencies can be seen in Figure 100 and Figure 101 in the Appendix):

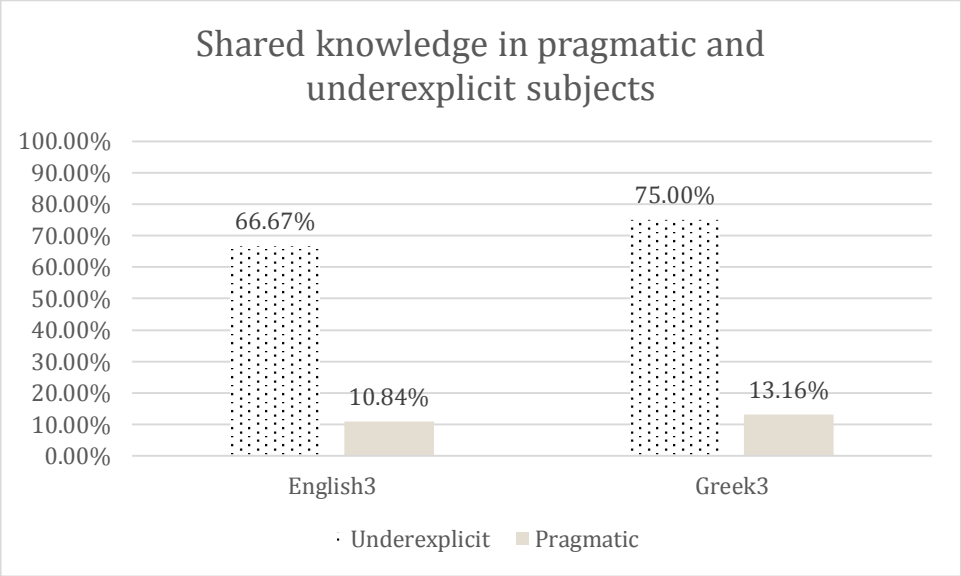


Figure 43. Shared knowledge factor in upper-advanced learners

Two post-hoc pairwise comparisons with Fisher's exact tests were performed (one for each group) between the frequency of underexplicit subjects which are constrained by shared knowledge and the frequency of the same constraint in pragmatic subjects. The results for both learner groups are highly significant (for the English3 group: $p<.0001$ and for the Greek3 group: $p<.001$). Recall here that the same comparison showed also highly significant differences for the native group (see section 6.3.1.4). Thus, these results further corroborate the hypothesis that upper-advanced learners tend to produce native-like anaphoric discourse in terms of the complex interaction between pragmatic discourse factors and the referential choices that they make.

In sum, regarding underexplicit referential discourse, as defined for the aims of this study, the native speakers have been found to be significantly more ambiguous than the lower proficiency learner groups (intermediate and advanced). On the other side, the highest proficiency levels of both English-speaking and Greek-speaking learner groups approximate the linguistic behaviour of the native speakers. Overall, the imperfections in the production of the learner groups reflect a more conservative referential strategy which is also in line with the results of the analysis of overexplicitness. The results broadly

confirm Hypothesis III (a) and run against Hypothesis III (b). It should be noted, however, that the highest proficiency learners are less bound up in a tendency to avoid ambiguity which for the less proficient groups (probably in conjunction with some cross-linguistic influence) may result in overexplicit production. At best (as in the case of the Greek3 group), learners may avoid unnaturally redundant discourse but, overall, they barely reach native-like usage of refined minimalist strategies in what concerns their referential choices. The implications of these findings, in relation to the previous literature, shall be broadly considered in the general discussion.

In the upcoming section, a full developmental account for the learner groups will be provided. The three proficiency groups of the English and Greek learners of Spanish will be contrasted to each other in order to examine how proficiency level may have an effect on the acquisition of anaphoric subjects.

6.3.2 English and Greek learners: developmental account

In this section, the results of the three proficiency levels of English-speaking and Greek-speaking learners will be considered in terms of developmental changes in their overexplicit and underexplicit anaphoric production. We start by considering the overexplicit production of the six English-speaking and Greek-speaking groups in Figure 44 (the original UAM CorpusTool raw frequencies can be seen in Figure 85 in the Appendix):

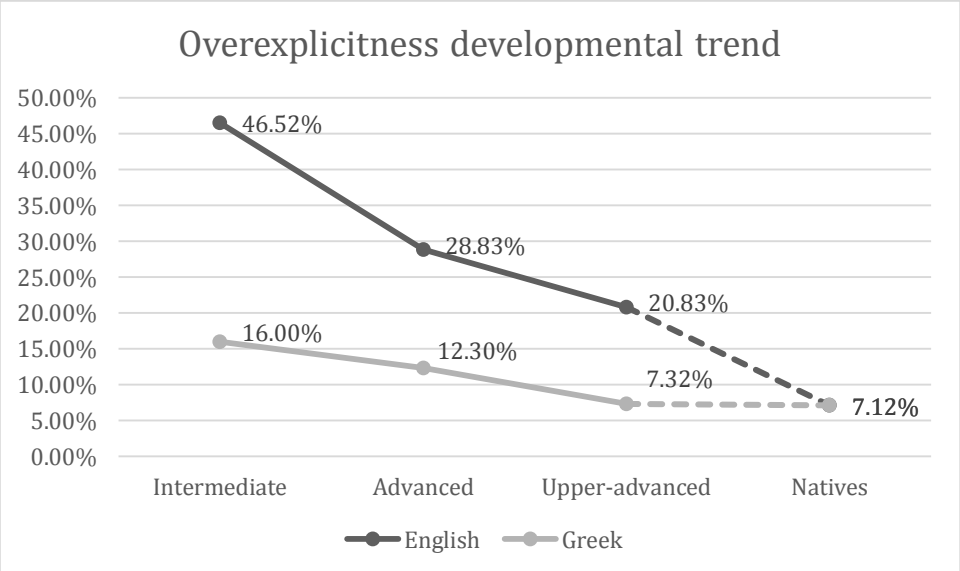


Figure 44. English learners: overexplicitness by proficiency level

Overall, the results indicate a clear developmental trend towards native-like performance in the anaphoric choices of both the English-speaking and the Greek-speaking groups.

Regarding the former, the upper-advanced English3 group is significantly less overexplicit than both the advanced English2 ($\chi2=5.31$, $p=.0212$) and the intermediate English1 group ($\chi2=52.78$, $p<.0001$). The advanced English2 group is also significantly less overexplicit than the intermediate English1 group ($\chi2=18.68$, $p<.0001$). The results, regarding overexplicitness, indicate that the most proficient English-speaking learners are performing in a more native-like way than the less advanced learners. The results indicate a clear developmental trend in the anaphoric production of the Greek-speaking groups as well. The statistical comparisons (Fisher's exact tests) revealed that the upper-advanced Greek3 group is significantly less overexplicit than the intermediate Greek1 group ($p=.0079$), whereas the difference between the upper-advanced Greek3 and the advanced Greek2 group only approximates but does not reach significance ($p=.0849$). Additionally, there is no significant difference between the overexplicit production of the intermediate Greek1 and the advanced Greek2 group ($p=.3345$). The results indicate that the differences, regarding the Greek-speaking learners, are less pronounced than in the case of the English-speaking groups. This is an expected result since the Greek learners are producing, overall, much fewer redundant subjects. It is striking that even the lowest proficiency intermediate Greek1 group are is overall less overexplicit (16%) than the highest proficiency upper-advanced English3 group (20.83%). Recall here that the participants of the former group scored less than 70% in the proficiency test, whereas the participants of the latter scored more than 95%.

The distributions of overexplicit production by type for the English-speaking and Greek-speaking learner groups was further examined in order to fully account for the potential causes of this phenomenon. In Figure 45 we see how the overexplicit anaphoric subjects are distributed by type for each group (the original UAM CorpusTool raw frequencies can be seen in Figure 85 in the Appendix):

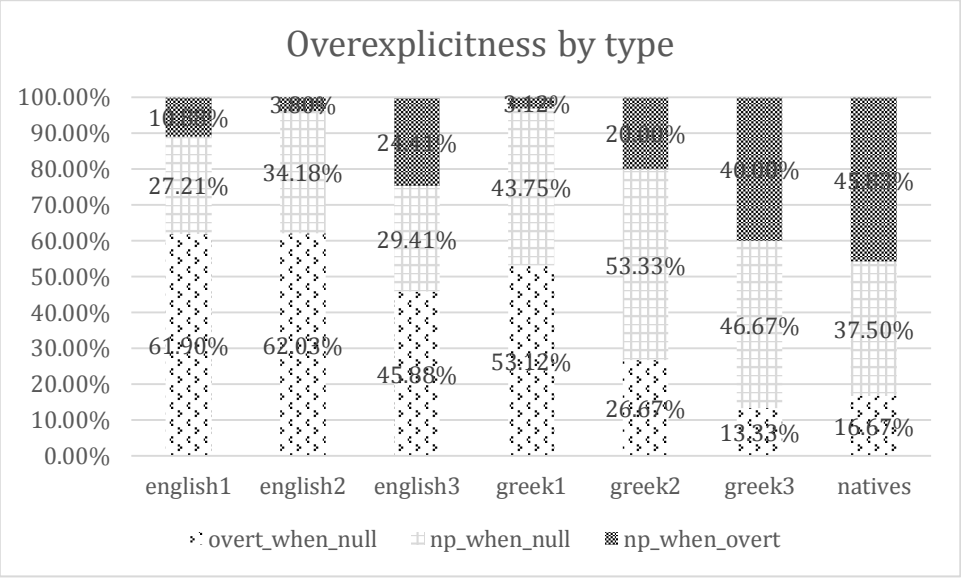


Figure 45. English learners: overexplicitness by type

As can be observed, the patterns of overexplicit production are very similar between the three English-speaking groups. Crucially, overexplicitness in the production of the English-speaking learners primarily concerns the use of redundant overt pronouns which, for all groups, constitute the most frequent overexplicit subject form. Note, however, that the upper-advanced English3 group is producing significantly less redundant 'overt_when_null' pronominals than both less proficient groups (for English1: $\chi2=4.98$, p=.0256 and for English2: $\chi2=3.67$, p=.0428). At the same time, the English3 participants produce significantly more 'np_when_overt' subjects than the English1 ($\chi2=6.68$, p=.0097) and the English2 learners ($\chi2=12.7$, p=.0003). These results further confirm a developmental trend towards the native patterns which, as we have already seen, mostly concern the production of redundant noun phrases. Regarding the Greek-speaking groups, we notice that, as proficiency grows, they produce less redundant pronominal subjects and more overexplicit noun phrases. This is the same tendency that was observed for the English-speaking groups. Note, however, that whereas the English learners mostly produce redundant pronominals at all proficiency levels, this is true only for the intermediate Greek1 group. The overexplicit production of the other two Greek groups mostly concerns redundant noun phrases. Regarding the 'overt_when_null' type, the difference between the production of the Greek1 group and the other two groups is significant (for the Greek2 (Fisher's exact test): p=.0415 and for the Greek3 (Fisher's exact test): p=.0118). In contrast, the Greek3 group produces significantly more 'np_when_overt' subjects than the other two groups (for the Greek2 (Fisher's exact test): p=.0412 and for the Greek1 (Fisher's exact test): p=.0026). Overall, these results confirm

the developmental trend towards the native speakers' anaphoric production which, regarding overexplicitness, mostly comprises noun phrases. Additionally, regarding the type of overexplicitness for each proficiency level, only the intermediate Greek-speaking learners behave similarly to the English-speaking groups. This might indicate some L3 influence for the lowest proficiency learners (recall here that all Greek learners have some linguistic competence in English as well). The advanced and upper-advanced groups are more similar to the native speakers, regarding the observed patterns of overexplicitness.

Regarding underexplicitness, the results for the English-speaking and the Greek-speaking learner groups are graphically represented in Figure 46 (the original UAM CorpusTool raw frequencies can be seen in Figure 99 in the Appendix):

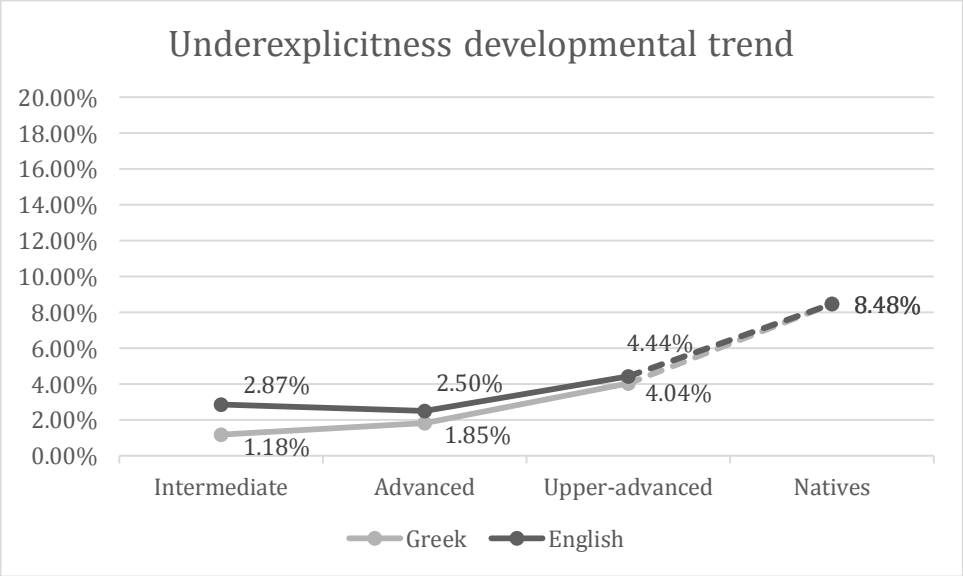


Figure 46. English learners: underexplicitness by proficiency level

The statistical comparisons (Fisher's exact tests) revealed that the differences regarding underexplicit subject forms between the three proficiency levels of the English-speaking groups are not significant (intermediate vs advanced: p=1, advanced vs upper-advanced: p=.3466, intermediate vs upper-advanced: p=.4757). However, a slight trend towards more native-like production as proficiency grows can be observed, notably in the upper-advanced group. Recall here that the native speakers were found to be significantly more underexplicit than intermediate and advanced learner groups, whereas the difference with the upper-advanced group was only marginally significant (see discussion in 6.3.1.6). Regarding the Greek-speaking learners, the same tendency observed for the English groups is also found here: while proficiency grows, the Greek learners are becoming more native-like regarding the production of underexplicit subjects. However, the differences

between the three proficiency levels of the Greek-speaking groups are not significant (intermediate vs advanced: p=.6984, advanced vs upper-advanced: p=.2441, intermediate vs upper-advanced: p=.1147).

In sum, despite the differences between the English and Greek-speaking learners, the participants of both L1 backgrounds are moving towards the native anaphoric patterns as proficiency grows, regarding overexplicit and underexplicit subjects. This finding supports Hypothesis IV regarding developmental improvement. Crucially, though, only the upper-advanced Greek learners become fully native-like in terms of overexplicit production whereas no strong claims can be made for underexplicitness (note, however, that statistical differences with the natives are marginally not significant). Overall, the results support Hypothesis II (a), III (a), IV, V and VI and run against Hypothesis II (b) and III (b). In sum, both English-speaking and Greek-speaking learners of Spanish present deficits and they both perform better as proficiency grows. Deficits mostly concern overexplicitness and, to a much lesser degree, underexplicitness (but not to the expected direction). The results regarding overexplicitness reveal that the role of crosslinguistic influence is crucial. However, other factors need to be considered as well in order to account for this extremely complex phenomenon. The results will be further discussed and compared to the findings of previous literature in the general discussion that follows.

CHAPTER 7

7 GENERAL DISCUSSION

The results presented in the previous chapter will be broadly discussed here in the light of the research questions and the hypotheses of this study (see Chapter 4), along with the findings and the claims made in previous theoretical and empirical literature on 3rd person anaphoric subjects (see Chapter 2 and Chapter 3). The first section of the general discussion deals with anaphora in Spanish L1 and the second section is dedicated to the main focus of this thesis, namely the acquisition of anaphoric subjects in Spanish L2.

7.1 Anaphoric subjects in Spanish L1

As we saw in Chapter 2 there are three fundamental assumptions made in the theoretical literature on discourse anaphora. First, several factors of both discursive and syntactic nature affect the dynamically-changing information status of a referent in discourse. Second, the aforementioned information status (expressed in terms of topicality/accessibility/givenness/activation/etc.) crucially determines the selection of anaphoric forms. Third, and closely related to the other two, less explicit forms (e.g. null instead of overt subjects) correspond to referents with higher information status (i.e. more topical/accessible/etc.) and vice versa. Overall, the results of this thesis broadly confirm the validity of the aforementioned assumptions for anaphoric subjects in Spanish L1. More specifically, the present study provides some answers to research question (1):

1. **Regarding the production of 3rd person anaphoric subjects in Spanish L1, what factors may account for the referential choices of the native Spanish speakers?**

The importance of two purely discursive factors, namely referential distance and interference, has been constantly highlighted in both theoretical and empirical studies on discourse anaphora (Ariel, 1990; Arnold, 1998; Givón, 1983; Toole, 1996; *inter alia*). The importance of referential distance was confirmed in the present study, insofar as it was found to directly determine the referential choices of the native speakers of Spanish: more explicit forms are produced as the distance from the antecedent grows. In terms of information status, when the antecedent is located far away from the anaphor, it is reasonable to assume that it will be less activated (Givón, 1983; Kibrik, 2011). Therefore, more explicit referential forms (overt subjects) are needed in order to unambiguously refer to it. Similarly, regarding PRI, the presence of more than one referent, as well as the gender of potential antecedents in the preceding discourse, was found to affect the

production of anaphoric subjects: more explicit forms are produced as the number of active referents in the preceding discourse grows, especially when the potential antecedents are of the same gender (Fox, 1987a; Lozano, 2016). In terms of information status, it is reasonable to assume that the presence of competitors may abase the information status of the referent and render it less accessible (Ariel, 1996). Therefore, more explicit forms are needed in order to unambiguously refer to it. It was further confirmed that the starting of a new paragraph constitutes a barrier for null anaphoric subjects (Hinds, 1977; Hofmann, 1989; Lozano, 2016) and promotes instead the production of overt forms. In other words, the evidence suggests that the information status of the referent tends to be 'zeroed' at the beginning of a new discourse segment (new paragraph), where overt subjects are massively produced. The effect of priming, previously suggested in the literature to constrain referential choices (Cameron, 1994; Travis, 2005, 2007) was found to be less decisive than the three aforementioned discourse factors. Finally, neither the protagonist status of the referent (Givón, 1990; Kibrik, 2000) nor the presence of shared knowledge (Blackwell, 1998; Prince, 1981a) were found to significantly affect the information status of the referent. Notice, however, that these two factors have been proposed in the theoretical literature on discourse anaphora but they have not been empirically tested before in Spanish. Therefore, more research is needed regarding their relevance in the production of anaphoric subjects.

Turning our attention from discursive to syntactic constraints, the results of this study confirmed that switch-reference is one of the main factors that determine referential choices in Spanish L1 (Abreu, 2009; Bentivoglio, 1983; Cameron, 1994; Cameron & Flores-Ferrán, 2004; Flores-Ferrán, 2010; Geeslin & Gudmestad, 2016; Otheguy & Zentella, 2012; Shin & Cairns, 2012; Shin & Otheguy, 2009; Silva Corvalán, 1982, 1994). More specifically, when the antecedent is located in the subject position of the exact previous clause, there is a very strong preference for the production of null anaphoric subjects. In terms of information status, the subject position seems to entail an a priori higher activation status for the referent, which is further reinforced by the closer distance in the case of same-reference patterns. This finding is also broadly in line with the effect

of the topic-shift feature, proposed under formal/generative accounts[127] (Lozano, 2009b, 2016; Sorace et al., 2009; Zulaica-Hernández, 2016; *inter alia*). Additionally, the privileged role of subject antecedents was further confirmed when the relevance of the syntactic function was examined independently of distance (Antecedent Syntactic Function factor). Consider the following example of a subject antecedent which is not located in the previous clause (i.e. a switch-reference with the antecedent in subject position):

158) Así $Ø_i$ se ve envuelto en un juego de ruleta rusa macabro, en el que una serie de señores$_j$ apuestan grandes cantidades de dinero. Por fortuna **$Ø_i$** logra ganar.

"So (he)$_i$ gets involved in a macabre Russian roulette game in which some men$_j$ are betting big amounts of money. Luckily, **(he)$_i$** manages to win."

The results indicate that more null anaphors are produced when the antecedent is in subject position, irrespectively of its location in the previous clause or not. This is fully in line with the traditional claim in the literature regarding the overall prominence of subject position (Chafe, 1976; Hobbs, 1979; Kaiser & Trueswell, 2008; Miltsakaki, 2002, 2007; Rohde & Kehler, 2014; *inter alia*) as well as with the claims made under the PAS account (Alonso-Ovalle et al., 2002; Carminati, 2002; Keating et al., 2011; *inter alia*). Finally, the type of clause (main, subordinate, coordinate) was also found to crucially affect the referential choices of native speakers of Spanish. The results of this study indicate that the information status of the referent (and consequently the use of more or less explicit forms) is highly reliant on the type of clause. Crucially, more explicit forms are produced in independent main clauses (i.e. after a full stop) than in dependent subordinate and coordinate clauses. Crucially, the same analysis as in the case of the new paragraph barrier may apply here. It seems that referents are more prone to keep their activation status in subordinate and coordinate clauses than in main clauses, where the presence of a full stop may act similarly (though not to the same degree) to the starting

[127] Note, however, that switch-reference and topic-shift are similar but not identical (see also section 5.5.2.1). Crucially, switch-reference does not entail a categorical association with overt pronouns, in contrast to topic-shift (Sorace & Filiaci, 2006; Tsimpli, Sorace, Heycock, & Filiaci, 2004). In many cases, when other factors boost the activation of a referent, pragmatically correct null subjects may be used in switch-reference contexts (see also section 2.4.1).

of a new paragraph. Additionally, coordinate clauses seem to foster the production of null subjects even more than subordinate clauses. This is reasonable, given that the former type may be assumed to entail the more structurally continuous relation among the three clause types. Consider the following examples:

159) (a) María$_i$ se levanta. **Ø$_i$** Tiene que ir a trabajar.
"Mary$_i$ gets up. (**She**)$_i$ has to go to work"
(b) María$_i$ se levanta porque **Ø$_i$** tiene que ir a trabajar.
"Mary$_i$ gets up because **(she)$_i$** has to go to work"
(c) María$_i$ se levanta y **Ø$_i$** tiene que ir a trabajar.
"Mary$_i$ gets up and **(she)$_i$** has to go to work"

Note that the anaphoric subject of the main clause in (159a) is produced after a full stop which may slightly decrease the activation of the referent (similarly to the starting of a new paragraph). In contrast, in the second version of (roughly) the same proposition in (159b), the anaphoric subject is produced in a dependent clause where no activation barrier is present. Same as in (159b), the anaphoric subject in (159c) is produced in absence of any structural barriers. In addition, the presence of the cumulative conjunction "and" may indicate the higher degree of continuation among the three types (as opposed to the full stop in the main clause or the subordinate conjunction "because"). Note, additionally, that the coordinate clause is the only of the above structures that allows for a null subject in both English and Spanish. Furthermore, although the presence of an overt pronoun (instead of null) in all three Spanish clauses is redundant, it may be the case that not all of them are equally inappropriate. As a matter of fact, an overt pronoun in the coordinate clause in (159c) would be extremely odd in Spanish (but not equally odd in the other two clause types). Although the presence of more null subjects in coordinated clauses in English and Spanish has been pointed out in previous literature (Nariyama, 2004; Shin & Montes-Alcalá, 2014; *inter alia*), no explanation (in discursive terms) has been given so far regarding this phenomenon. We believe that the novel observations made here merit to be empirically examined in the future, ideally in experimental settings.

In sum, we found that several discursive and syntactic factors affect the information status of the referent and, consequently, the production of anaphoric subjects in Spanish L1. More specifically, the main factors that were found to be relevant in Spanish L1 are: Distance, PRI, New Paragraph, Switch-Reference, Antecedent Syntactic Function and Clause Type. This is fully in line with the theoretical literature on anaphora and other previous empirical studies on anaphoric subjects in Spanish L1. However, as Arnold (2003:226) argues, "the job for researchers of language processing, language production,

pragmatics, and computational linguistics is to determine what these factors are, how they interact, which are the most important, and how they contribute to our ability to use language". Regarding Spanish L1, this study has only addressed the first of these issues. More specifically, the methodological approach adopted in this study did not allow us to address two crucial questions. First, the particular relevance of each factor in presence of all other factors, i.e. the interaction of each factor with all other factors. Second, the potential fine-grained differences between the Spanish overt subjects forms (overt pronouns, several types of noun phrases, demonstratives, etc.). Note, however, that the consideration of both aforementioned issues requires the strict cooperation with a statistician for the design of a sophisticated multifactorial regression model with a non-binary dependent variable (null, overt pronoun, NP, etc.) and several interacting factors of different nature (categorical, discrete, continuous, etc.). The only study, to my knowledge, that has employed such a model is Gudmestad et al. (2013)[128]. Despite the rigors of this approach, we strongly believe that future research on discourse anaphora in Spanish L1 should benefit from the application of sophisticated statistical models in large corpora (Gries, 2015; Gries & Deshors, 2014).

7.2 Anaphoric subjects in Spanish L2

Regarding the acquisition of 3rd person anaphoric subjects in Spanish L2, a fine-grained approach was adopted for the analysis of the English-speaking and Greek-speaking learners' data (see section 6.3). Focusing on the pragmaticality of anaphoric subjects, we objectively operationalized and defined overexplicitness and underexplicitness according to specific discursive and syntactic criteria (see sections 2.4.3 and 5.5.3). Then we sought to provide some answers to several research questions that have been addressed in previous SLA literature, namely:

[128] Recall that Gudmestad and colleagues focused on anaphora in Spanish L2. We believe, however, that their model would be more adequately applied to Spanish L1. As already argued (see sections 3.3 and 6.3), the direct application of anaphora models on L2 discourse may not reveal the fine-grained differences between the anaphoric production of L2 learners and native speakers.

2. (a) Do learners of Spanish from both L1 backgrounds (English and Greek) show deficits in the production of 3rd person anaphoric subjects?

(b) May the properties of 3rd person anaphoric subjects in Spanish L2 be eventually acquired?

There is a considerable amount of evidence in SLA literature pointing to the fact that L2 learners of null-subject languages such as Spanish and Italian have important difficulties with the interpretation and distribution of 3rd person anaphoric subjects (Ballester, 2013; Bel & García-Alcaraz, 2015; Bel, Sagarra, Comínguez, & García-Alcaraz, 2016; Belletti et al., 2007; Bini, 1993; García-Alcaraz & Bel, 2011; Jegerski et al., 2011; Keating et al., 2011; Lozano, 2002, 2009b, 2016, forthcoming; Margaza & Bel, 2006; Montrul & Rodríguez Louro, 2006; Pérez-Leroux & Glass, 1999; Rothman, 2007, 2009; Sorace & Filiaci, 2006). Although the bulk of previous evidence concerns learners with English L1 background, some of the aforementioned studies have demonstrated that learners from other L1 backgrounds (e.g. Arabic, Greek, Farsi and Italian) show deficits as well. In line with the previous literature, the results of the present study indicate that both English-speaking and Greek-speaking learners of Spanish L2 experience difficulties with anaphoric subjects. More specifically, they produce pragmatically inappropriate anaphoric subjects to a higher degree than native speakers of Spanish. It should be noted that this finding, which provides a positive answer to research question (2a), was broadly overlooked during the first years of SLA studies on the acquisition of anaphora due to the complexity of the phenomenon, reflected in the subtle difference between 'correct' and 'right' anaphoric forms. As Huang (2000b) points out:

> For any entity to which reference is to be made in discourse, there is a (potentially large) set of possible anaphoric expressions each of which, by a correspondence test, is 'correct' and therefore could in principle be used to designate that entity. On any actual occasion of use, however, it is not the case that just any member of that set is 'right'.

In line with more recent approaches, it is crucial to realize that all referential forms are grammatically correct in discourse and only their fine-grained examination in terms of pragmatic appropriateness may reveal whether L2 learners experience difficulties with their use and interpretation. Recall that the focus in current SLA research has shifted towards the consideration of the discursive conditions under which learners and native speakers produce anaphoric forms. This has led some scholars to propose that phenomena such as anaphora, located at the interface between syntax and discourse, are inherently and insurmountably problematic (Sorace, 2011; Sorace & Serratrice, 2009; Valenzuela,

2006). However, the present study provides evidence against the aforementioned hypothesis (the strong version of the Interface Hypothesis), insofar as the upper-advanced Greek-speaking learners of Spanish were found to perform native-like regarding the production of 3rd person anaphoric subjects (a prototypical syntax-discourse interface phenomenon). This finding is in line with the results of other recent studies on the acquisition of anaphoric subjects (Bel, Sagarra, Comínguez, & García-Alcaraz, 2016; Ivanov, 2012; Judy, 2015; Kras, 2008; Rothman, 2009; Slabakova, Kempchinsky, & Rothman, 2012; Zhao, 2014) and demonstrates that "the syntax-pragmatics interface is not a predetermined locus of fossilization" (Rothman, 2009:951). The present study, in line with the aforementioned studies, provides evidence that some linguistic features located at the interface between syntax and discourse may be fully acquired. However, this does not entail that all features related to the syntax-discourse interface will be unproblematic. As a matter of fact, an important number of studies have demonstrated that some interface-related features are difficult to acquire even at very advanced levels of proficiency (Ballester, 2013; Belletti et al., 2007; Lozano, forthcoming; Sorace & Filiaci, 2006; *inter alia*). All in all, the present study does not aim to test the original ambiguous version of the IH (Sorace, 2011:25), which predicts that interface-related properties "may not be fully acquirable" (Sorace & Filiaci, 2006:340), because "the use of 'may not' and 'fully' covers every possible situation in the acquisition of external interfaces making IH unfalsifiable" (Zhao, 2014:383). The results of this study merely indicate that some syntax-discourse properties can be eventually acquired (Lozano, forthcoming; Montrul, 2011; Rothman, 2009; White, 2011). In sum, regarding research question (2) and the corresponding hypotheses, learners were found to show deficits in the production of 3rd person anaphoric subjects but we also found that these deficits may be overcome, as the data of the upper-advanced Greek-speaking learner group demonstrate.

3. If non-target L2 performance is observed, does it concern overexplicitness, underexplicitness or both?

Research question (3) was aimed to test the UDH (Sorace, 2004, 2006a) according to which only overuse of overt subjects (overexplicitness) and no overuse of null subjects (underexplicitness) should be expected in L2 discourse. More specifically, this hypothesis predicts that L2 learners will produce redundant anaphoric subjects to a higher degree than the native speakers but, crucially, they will not overproduce ambiguous null subjects. The results of this study are in line with this hypothesis. On the one hand, learners produce

a considerable amount of overexplicit anaphoric subjects. This is an expected finding given that overexplicitness in L2, notably at intermediate/advanced proficiency levels, has been widely claimed to be a universal phenomenon (Ryan, 2015:824). There is a considerable amount of evidence in the SLA literature of the last three decades (apart from the numerous 'interface studies') that point to this direction (Chini, 2005, 2009; Fakhri, 1989; Gundel, Stenson, & Tarone, 1984; Henriëtte Hendriks, 2003; Kang, 2004; Leclercq & Lenart, 2013; Polio, 1995; Ryan, 2015; Williams, 1988). The present study confirms the observations of Hendriks (2003:292) and Ryan (2015:827) who argue that overexplicitness is a general feature of L2 discourse mostly associated with intermediate levels of proficiency, irrespective of source and target language. On the other hand, learners were found to produce very few underexplicit subjects. This runs against the findings of some studies that have reported both over- and underexplicitness in L2 (Lozano, 2009b; Montrul & Rodríguez Louro, 2006; Rothman, 2009). As a matter of fact, the L2 participants of the present study were found to produce less underexplicit subjects than the native speakers. This finding may seem counterintuitive at first. Note, however, that when redundancy and ambiguity are objectively defined according to specific criteria (in terms of discursive and syntactic factors) it is reasonable to expect that native speakers may also be redundant and/or ambiguous (given that they are not the gold standard under which inappropriateness is determined). As a matter of fact, previous literature confirms that native speakers are occasionally redundant (Alonso-Ovalle, Fernández-Solera, Frazier, & Clifton, 2002; Bel, Perera, & Salas, 2010; Blackwell, 1998; Lozano, forthcoming; Perales & Portillo, 2007). Additionally, as Abreu (2009:23) notes "usage-based research on NS production of SPPs (subject personal pronouns) in Spanish (…) has shown that NSs frequently produce SPPs that are technically redundant". Regarding underexplicitness, the results of the present study are in line with Blackwell & Quesada (2012) who also found that native speakers of Spanish choose a significant amount of ambiguous null subjects (12%) that may lead to a breakdown in communication. The authors concluded that "in every case, NSs are more likely than learners to use a 'less specific' form". The results of the native speakers are also broadly in line with the 'avoid pronoun' principle (Chomsky, 1981) that dictates the use of null subjects unless impossible (Rothman, 2009:955). Crucially, it was revealed that the underexplicit production of native speakers does not lead to insolvable ambiguity owing to the presence of shared information (previous discourse and/or world knowledge). In other words, native speakers' discourse seems to incorporate pragmatic considerations of economy to a higher degree than learners' production which is mostly characterized by redundancy.

As it has been argued (see section 6.3.1.4), the very limited underexplicit production of L2ers is consistent with their overall tendency to be overexplicit. In other words, learners' deficits mainly concern overexplicitness, insofar as their target-deviant underexplicit production may be accounted for as an epiphenomenon of overexplicitness. In sum, in line with the PPVH in Lozano (2016) as well as with the UDH in Sorace (2004, 2006), L2 learners were found to be more redundant than ambiguous.

4. Is L2 performance affected by proficiency level?

The results of the present study suggest that L2 performance, with respect to the production of anaphoric subjects, is affected by the proficiency level of learners. This is in line with the bulk of previous literature on the acquisition of anaphoric subjects, where learners of Spanish have been found to perform better as proficiency grows (Ballester, 2013; Geeslin, Linford, & Fafulas, 2015; Lozano, forthcoming; Margaza & Bel, 2006; Montrul & Rodríguez Louro, 2006; Rothman, 2009). The results of the present study converge with previous studies where more than one proficiency level has been jointly considered and compared. In sum, there is compelling evidence in the literature regarding the relevance of the 'proficiency level' factor. In addition, the results of two recent SLA studies are fully in line with the developmental account provided in the present study. Regarding the development of English-speaking learners of Spanish in L2 corpus-based studies, Geeslin et al. (2015) tested five levels of proficiency and found a U-shaped behaviour in the production of overt anaphoric subjects. Crucially, the intermediate English-speaking learners showed the highest rates of production of overt subjects, whereas production rates linearly decreased in advanced and upper-advanced groups. This is in line with the results of the English-speaking groups of the present study. Regarding the development of Greek-speaking learners, Margaza & Bel (2006) tested two Greek-speaking groups of different proficiency levels (intermediate and advanced) and found that the former was overexplicit whereas the latter did not overuse pronominal subjects. This is also in line with the results of the present study with respect to the Greek-speaking groups.

Overall, there is solid evidence in the literature that, regarding anaphoric subjects, proficiency level crucially affects L2 performance. As already discussed (see section 5.2), this further highlights the need for independent proficiency tests in order to strictly control this variable and allow comparability between learner groups and between different studies.

5. Is the L1 a facilitating factor in the acquisition of anaphoric subjects in Spanish L2?

One of the main purposes of this study was to investigate the role of cross-linguistic influence in the acquisition of anaphoric subjects in Spanish L2. In order to do so we followed the premises of Myles (2015:315) who argues that "to understand whether a specific developmental pattern is due to transfer or to the inherent characteristics of the property to be acquired, a wide range of well-chosen L1s and L2s which exhibit different realizations of the property in question are required". Given that Spanish is a pro-drop language (see section 3.1), the methodological setting employed here may be considered ideal for this purpose: learners from a pro-drop L1 (Greek) were compared to learners from a non-pro-drop L1 (English) at three objectively determined proficiency levels (see section 5.2). It was predicted that English-speaking learners would produce pragmatically redundant overt subjects to a greater extent than Greek-speaking leaners, due to the influence of their L1 (recall that overt subjects are obligatory in English). This was fully confirmed: crucially, the Greek-speaking learners performed significantly better than the English-speaking learners at all three proficiency levels. This provides solid evidence in favour of the role of cross-linguistic influence as a crucial factor in the acquisition of anaphoric subjects, in line with some previous studies (Bel, Sagarra, Comínguez, & García-Alcaraz, 2016; Belletti, Bennati, & Sorace, 2007; Fernandez de Moya, 1996; Gaillat, 2016; García-Alcaraz, 2015; Gundel & Tarone, 1992; Kang, 2004; Kras, 2008; Lozano, 2002b, 2002c; Prentza, 2014a, 2010; Slabakova & García Mayo, 2015; Tsimpli & Dimitrakopoulou, 2007; White, 1985, 1986) and against others (Bini, 1995; Chini, 2005, 2009; Lozano, forthcoming; Polio, 1995; Sorace & Filiaci, 2006; Sorace & Serratrice, 2009; Valenzuela, 2006).

The facilitating role of the L1 is further confirmed through a rather novel finding: although English-speaking learners massively produce redundant overt subjects, they do so to a significantly lesser degree in same-subject coordination patterns. Crucially, these are the only structures (among the different structures examined in the present study) where null subjects are allowed in English (Nariyama, 2004:240). The results regarding coordination confirm an observation made in Montrul & Rodríguez Louro (2006) according to which English-speaking learners of Spanish produce null subjects mostly in

coordinated structures[129]. The authors noticed that "since English allows null subjects in these situations (Haegeman, 1997) this result is not surprising" (p.412). Finally, the facilitating role of the L1 is clearly reflected in the fact that English-speaking learners need significantly more instruction time (nearly the double) than the Greek-speaking learners to achieve similar scores in the same independent proficiency test (see section 5.2).

Overall, the evidence provided in the present study highlights the role of the L1 as a crucial factor (although not necessarily the only one) in the acquisition of anaphoric subjects. The results broadly confirm the claims made in Gass & Selinker (1992) who early on pointed out that "if the L1 and the L2 share a parameter setting, this might be expected to offer an advantage to the language learner, and lead to some kind of positive transfer". This finding is broadly in line with some representational accounts (UA and IPH, Tsimpli & Dimitrakopoulou, 2007; Tsimpli et al., 2004) and runs against the processing accounts related to the IH (Sorace & Filiaci, 2006; Sorace & Serratrice, 2009). Crucially, the former highlight the role of cross-linguistic influence whereas the latter promote, instead, a processing explanation for L2 non-native performance. The reader is referred to section 3.2.4 for more details on these accounts.

In sum, this study has provided solid evidence for the role of transfer in the acquisition of anaphoric subjects in Spanish L2 which "can only be convincingly shown to occur when learners of different L1s learning the same L2 behave differently, and this difference in behaviour can be traced back to the L1" (Myles, 2015:317). Our results lead us to agree with Prentza (2014b:382) who argues that "in the examination of L2 pronominal use and interpretation the factor of cross-linguistic influence not only is relevant but, probably, should be considered first, over factors which have to do with the interface status of the phenomenon". That being said, our results also lead us to consider other factors that may act in conjunction with transfer and will be examined in the answer to the next research question.

[129] Note that Montrul & Rodríguez Louro (2006) merely observed that less redundant subjects were produced in coordination structures but, crucially, did not report any statistical tests to confirm this. The present study is the first to empirically confirm that English-speaking learners of L2 Spanish produce significantly less redundant anaphoric subjects in coordinate clauses.

6. If non-target L2 performance is observed, what factors may account for it?

The results of the present study revealed that cross-linguistic influence crucially affects the acquisition of anaphoric subjects in Spanish L2. However, a critical question remains: is transfer the only factor that accounts for non-target L2 performance? The evidence of the present study, in line with other previous literature (Bini, 1993; Chini, 2005, 2009; Lozano, forthcoming; Margaza & Bel, 2006; Polio, 1995), clearly suggests that the picture is far more complex than that. If Greek-speaking learners had been found to perform native-like at all proficiency levels, we could have solidly concluded that the problematic performance of the English-speaking groups is only due to negative cross-linguistic influence. However, some Greek-speaking groups (notably the intermediate and, to some extent, the advanced) were found to produce overexplicit anaphoric subjects in contexts where their L1 (Greek) dictates the selection of less explicit forms (same as in Spanish). This linguistic behaviour arises two important questions. First, what factors may account for the non-target performance of the Greek-speaking learners? Second, may the same factors be active (in conjunction with transfer) in the production of the English-speaking learners? In order to answer these questions, an overview of the various factors that have been proposed in the literature to account for the (non-)acquisition of anaphoric subjects is in order.

L2 transfer

To begin with, the role of previous learning experience has been considered in several 'L2 to L3 transfer' models (see Slabakova & García Mayo, 2015 for an overview). According to the L2 Status Factor model (Bardel & Falk, 2007; Falk & Bardel, 2011), the role of a previously learned L2 is stronger than the role of the L1 at the initial stages of the acquisition of morphosyntax. Given that the Greek-speaking learners of Spanish L2 have some previous knowledge of English due to the fact that this is studied since the first years of education in Greek primary schools (see section 5.2), it is reasonable to assume that there may be some influence from English in their production of anaphoric subjects. The following example from the intermediate Greek-speaking group provides some evidence in this direction:

160) Las letras de Manu Chao canciones hablar de amar (GR22_21_0_2_christos)

"The lyrics of Manu Chao's songs are about love"

Crucially, the genitive construction "de Manu Chao canciones" in example (160) is ungrammatical in both Spanish and Greek (the grammatically correct order would be

"canciones de Manu Chao"). This grammatical error can be straightforwardly explained as a word-by-word translation from the corresponding English construction "of Manu Chao's songs". Although this example does not concern the production of anaphoric subjects, it clearly demonstrates how previous knowledge of English may affect the production of Greek-speaking learners of Spanish[130]. In the same direction, Abreu (2009:61) reports the results of De Angelis (2005) who found similar effects for L2 learners of Italian and points out that "parallel language activation may be partly responsible for production or omission of subjects in a third or fourth language". All in all, more research is needed in this area. Ideally, in the future, Greek-speaking learners of Spanish with knowledge of English should be compared to learners without previous knowledge of English. As already argued, however, the task of collecting a participant pool with these characteristics is very difficult due to the strong presence of English language in the Greek educational system and society.

Input

Another factor that has been claimed to account for the acquisitional problems of L2 learners regarding anaphoric subjects is input (Keating, VanPatten, & Jegerski, 2011; Linford & Shin, 2013; Lozano, forthcoming; Pladevall-Ballester, 2013; Polio, 1995; Rothman, 2007, 2009). In this direction, it has been argued that in explicit instruction (materials and textbooks) the fine-grained differences between anaphoric forms in discourse are not treated (Pérez-Leroux & Glass, 1999:230). Additionally, some authors argue that L2ers may not be receiving fully native-like input during the first stages of acquisition (Rankin, 2015; Tono, 2004). More specifically, the idea is that instructors and native speakers may be employing emphatic speech in order to ensure successful communication with L2ers, to the detriment of pragmatic appropriateness in the use of anaphoric subjects. In other words, more explicit discourse than usual (i.e. overt subjects) may be used in the interaction with learners, especially at lower proficiency levels. Although this is a rather unexplored area, there are at least two studies that empirically confirm this hypothesis. Crossley, Louwerse, McCarthy, & McNamara (2007) examined

[130] Recall that during the annotation procedure, we focused exclusively on 3rd person anaphoric subjects. Given that no attention was given to cases of potential 'L2 to L3' influence, the example (160) was accidentally discovered. Therefore, we are not in a position to examine whether there are more similar cases in our data or not. Whatever the case, the aforementioned example clearly demonstrates that the knowledge of a previously learnt L2 may affect the performance in another foreign language.

simplified texts that are broadly used in textbooks for beginning and intermediate-level L2 learners of English and compared them to authentic language. The authors found that simplified texts show a significantly higher number of constituents than authentic texts and demonstrated that "simplified texts, in their effort to provide more accessible language, may depend too heavily on certain constructions, such as noun phrases" (p.26). The finding that simplified texts may contain atypical language structures is also in line with other previous studies (Kennedy & Bolitho, 1984; Willis, 1998). Additionally, Dracos (2010) reported very recently that, in classroom contexts, instructors of Spanish L2 produce higher rates of overt pronouns than normal. Finally, Lozano (forthcoming) argued that the attested optionality in the interpretation of anaphoric subjects by learners of L2 Spanish "may be a reflection of the ambiguous input they may be receiving". In sum, it can be concluded that there is some preliminary evidence pointing to the direction of 'unpragmatic input' as another factor that may affect the non-native performance of L2 learners. Given the potential relevance of this factor, more research is needed in this area.

Ambiguity avoidance

A third explanation that has been repeatedly offered in the literature has to do with an 'ambiguity avoidance' strategy that L2 learners may rely on due to universal principles of clarity (Bini, 1993; Chini, 2005; Fakhri, 1989; Hendriks, 2003; Keating et al., 2011; Leclercq & Lenart, 2013; Lozano, 2016; Polio, 1995; Rothman, 2009; Ryan, 2015; Shin & Cairns, 2009, 2012; Sorace & Filiaci, 2006; Williams, 1988). This explanation has been recently resumed under the Pragmatic Principles Violation Hypothesis (PPVH) proposed by Lozano (2016). According to the PPVH and other similar accounts, learners are redundant due a default strategy whose purpose is to avoid ambiguity and the corresponding communication breakdown. More specifically, "learners are aware of their "short-comings" in the L2" (Hendriks, 2003:294) and, thus, exhibit "a general tendency to err on the side of caution" (Shin & Cairns, 2009:162). In words of Rothman (2009:967): "After all, overusing overt pronouns is not wrong per se. It is simply pragmatically odd. Worse, however, is the failure to use overt subjects when the discourse information does not provide an immediately identifiable/accessible subject". In sum, learners might generally follow an 'avoid miscommunication' principle (Sorace & Filiaci, 2006:348) that may lead to redundancy whereas native speakers, as already discussed, mostly follow an 'avoid pronoun' principle (Chomsky, 1981) that may lead to ambiguity. This account would provide some explanations regarding the overexplicit

production of the Greek-speaking learners. At the same time, given the universality of such principles, it could be expanded to partly account for the production of English-speaking learners as well (in conjunction with transfer).

Processing difficulties

Several authors have argued in favour of processing explanations for the non-target performance of L2 learners (Arnold, 2010; Chini, 2005; Leclercq & Lenart, 2013; Linford & Shin, 2013; Ryan, 2015; Torregrossa & Bongartz, forthcoming; Vogels, Krahmer, & Maes, 2015). According to these accounts, overexplicitness is triggered when the control mechanisms of L2ers are being taxed due to the cognitive load involved in the selection of referential expressions in real time. This is also consistent with the processing accounts related to the IH (see section 3.2.4), according to which non-target L2 performance is due to the processing difficulties involved in the integration of linguistic knowledge at the interface between syntax and discourse (Sorace, 2011; Sorace & Filiaci, 2006; Sorace & Serratrice, 2009). The reader is referred to the overview provided in section 3.2.5 for more details.

Summary

In sum, at least five main factors have been proposed in the literature to account for the target-deviant performance of L2 learners: L1 transfer, L2 transfer, input, clarity principles and processing difficulties. To the aforementioned explanations we may add at least two others that, although less popular, have also been proposed by some authors. Hendriks (2003), following Véronique, Carroll, & von Stutterheim (2000), argues that overexplicitness in L2 discourse is due to the lack of the necessary linguistic means for reference maintenance and disambiguation. More specifically, L2ers may fail to create optimal contexts for the use of less explicit forms (e.g. null subjects in the case of Spanish). In words of the author: "The learner does not manage to construct one coherent whole. As a result, larger referential chains keeping the topic constant (topic persistency) will fail to occur, resulting in less optimal conditions for the use of pronominal forms" (p.295). On the other side, Bini (1993) and Polio (1995) argue that L2 learners overuse pronouns as a strategy that provides them more time to think. This might partly explain overexplicitness in oral discourse but it does not seem relevant for the present study that focuses on written essays.

In sum, a fairly complex picture emerges from all the aforementioned accounts taken together and, as very recently proposed by some authors, the need of integrating various

perspectives is made clear (Kibrik, 2011; Quesada, 2015; Ryan, 2015; Sorace, 2011, 2012; White, 2011). No single factor seems to be able to account for the non-target behavior of all learner groups together (irrespective of source-target language and proficiency level) (Ryan, 2015:852). In words of White (2011:588):

> There is no reason to assume that there is only one source of difficulty at the interfaces, either non-native grammatical representations or non-native processing. There may, indeed, be multiple explanations of L2ers' problems, including different sources for different interfaces, or for different linguistic phenomena, or for different levels of L2 proficiency.

Instead of adopting an 'either-or' perspective, a more sophisticated explanation of the results of this study requires to consider the interaction of multiple factors that may constrain the performance of L2 learners to different degrees according to proficiency level and L1 background (Sorace, 2011; Sorace, Serratrice, Filiaci, & Baldo, 2009). This multifactorial approach is graphically represented in Figure 47:

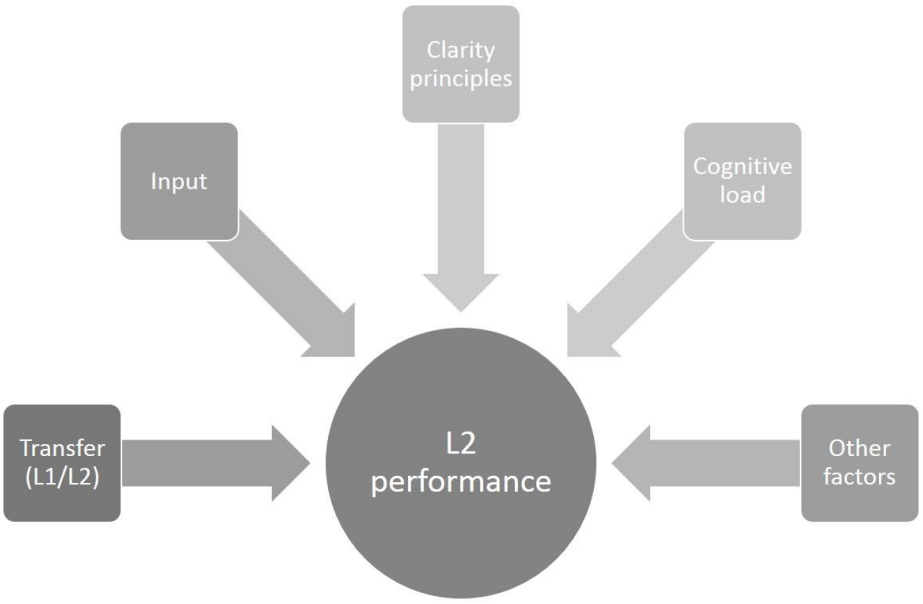


Figure 47. The interaction of multiple factors to L2 performance

As can be seen in Figure 47, several factors may cumulatively and simultaneously constrain L2 performance. The relevance of each factor may not be the same for all proficiency levels nor for all source/target language pairs. For example, the results of this study indicate a substantial L1 negative transfer effect for the English-speaking learners. This does not preclude that other confounding factors may be operative as well, resulting

in the massive production of overt pronouns observed for the intermediate English-speaking group. However, whereas the L1 effect may be more striking for lower-proficiency groups, other factors may be taking over as proficiency grows. Similarly, the Greek-speaking learners may be constrained by L2 English influence at lower levels of proficiency. Other factors may act in a cumulative way (e.g. input, clarity principles, cognitive load), resulting in the production of overexplicit subjects. Future research will need to determine the extent to which these interacting factors are operative for each source/target language pair and for each proficiency level. Finally, individual differences may also be at play in the production of anaphoric subjects. Although the methodological approach employed in the present study does not allow to consider this factor, we fully agree with Geeslin & Gudmestad (2011) who argue for the need to do so in future research. In words of the aforementioned authors: "We do not wish to imply that individual variation is of no interest. On the contrary, we hope to examine individual language use in greater detail once we have gained a better understanding of the group trends as a whole" (p.22).

CHAPTER 8

8 CONCLUSIONS

The conclusions reached and the overall contribution of the present study to current research on anaphora will be summarized in this section. Additionally, some limitations of this study will be pointed out and directions for future research will be suggested.

8.1 Summary of conclusions

The present study has focused on anaphora in Spanish L1/L2. Particular emphasis was given on the acquisition of 3rd person anaphoric subjects by learners of Spanish L2. For this purpose, this study is the first to examine groups of L2ers from two different L1 backgrounds (English and Greek) at three proficiency levels (intermediate, advanced, upper-advanced). Additionally, we tested the relevance of several factors that have been previously claimed to constrain the production of anaphoric forms in Spanish L1. The corpus methodology employed in this study allowed us to test a number of claims and hypotheses previously made in the literature regarding discourse anaphora and SLA. In this chapter, we present the conclusions that were reached and their implications for relevant research. Additionally, some of the limitations of the present study are discussed and some directions for future research are suggested.

The conclusions reached in the present study, in relation to the hypotheses, can be briefly summarized as follows:

I. The production of 3rd person anaphoric subjects in Spanish L1 can be properly accounted for as a result of the complex interaction between multiple syntactic and discursive factors.
II. Learners of Spanish L2 exhibit non-target performance (deficits) with some properties involved in the production of 3rd person anaphoric subjects. These deficits may be eventually overcome by some learners.
III. The major deficits observed in the performance of the L2 learners concern the production of overexplicit 3rd person anaphoric subjects.
IV. Proficiency level crucially affects the performance of L2 learners, insofar as more proficient groups perform better than less proficient groups.
V. The L1 is a crucial factor in the acquisition of 3rd person anaphoric subjects.
VI. The non-target performance of L2 learners may be better explained in terms of the influence of multiple factors.

It should be noted that the present study follows some recent research that highlights the need to take into consideration noun phrases in the analysis of 3rd person anaphoric subjects' production in Spanish L1/L2 (Gudmestad, House, & Geeslin, 2013; Lozano, 2009b, 2016). Despite the fact that NPs have been traditionally considered as one of the main referential forms in the theoretical literature on discourse anaphora (Ariel, 1990; Givón, 1983; Gundel, Hedberg, & Zacharski, 1993), they have been largely overlooked in early anaphora studies on Spanish L1/L2. Regarding Spanish L1, this might be associated with the fact that previous production-oriented research (mainly from a variationist perspective) has commonly supported the simultaneous consideration of all three grammatical persons. NPs may have been 'sacrificed' for comparability purposes, i.e. in order to allow the merge of 1st and 2nd persons (expressed exclusively with null and overt pronouns) with the 3rd person anaphors. However, this practice raises some critical issues of comparability (instead of resolving them). As already discussed (see section 2.3), 3rd person anaphoric subjects have fundamentally different anaphoric properties from the 1st and 2nd person and one of the main differences lies precisely in the fact that only the former may be expressed with a noun. Therefore, in line with Lozano (2009b) the present study argues for the need to examine grammatical persons separately and to include NPs in the analysis of 3rd person anaphoric subjects. Following some recent research (Blackwell & Quesada, 2012; Gudmestad, House, & Geeslin, 2013; Lozano, 2009b, 2016), this study tackles the aforementioned critical issues by focusing exclusively on 3rd person anaphoric subjects (including NPs). The strict control of grammatical person and the inclusion of NPs allowed the present study to contribute to a better understanding of discourse anaphora in Spanish L1 regarding 3rd person subject forms. More specifically, the results confirmed the relevance of some of the main discursive and syntactic factors proposed in the literature to account for the production of anaphoric forms in Spanish L1 discourse, insofar as more/less explicit forms (null/overt subjects) were found to be produced in relation to the presence/absence of these factors. However, as already discussed (see section 7.1), a limitation of our methodological approach lies in the fact that overt subjects (overt pronouns and the different types of NPs) were temporarily merged (only during the analysis of the Spanish L1 data) under a single 'overt subject' category. This did not allow us to consider potential fine-grained differences in the properties of the different overt subject forms and future research will need to address this question.

Regarding the acquisition of 3rd person anaphoric subjects by L2 learners, this study has contributed to some ongoing debates in SLA literature by examining this prototypical discourse-constrained phenomenon in real production data. As already discussed, the bulk of previous evidence in the L2 acquisition of anaphora is based on the results of offline experiments (see section 3.2) whereas the present study, in line with other recent research (Lozano, 2009b, 2016; Ryan, 2015), tested existing L2 acquisition hypotheses in real discourse. Given the nature of the data, however, particular emphasis was given to prevent potential comparability issues related with the complex characteristics of natural discourse. More specifically, the methodological setting employed here allowed us to test learners from two different L1 backgrounds at three different objectively-defined proficiency levels under the same conditions. This is crucial, given the complex picture that emerges from the findings of previous literature (see sections 3.2.6 and 3.3.3). As already discussed, the differences in the methodologies, the participant pools and the specific focus of interest of each study render unsafe the direct comparison of the reported results. Consequently, an important amount of contradictory evidence in previous literature could be attributed to methodological differences. As a result, despite the significant body of research in the last years, several crucial question remain unanswered. The present study is among the first to strictly control the comparability between the examined L2 groups by using an independent placement test to objectively categorize L2 learners per proficiency level (see section 5.2). Additionally, as another novelty of this study, we strictly controlled the comparability between the production of each group by establishing objectively defined discourse patterns of pragmatic and unpragmatic use (see sections 2.5 and 5.5.3). Finally, discourse genre and mode have been also controlled since the data of all groups consist of written narrative texts (see section 5.2). In sum, the present study is the first to compare the anaphoric production of seven different groups (six L2 groups and a native control group) under the exact same conditions.

Once the valid comparability between the participant groups and their written production was ensured, the inclusion of two different L2 populations (with respect to the L1) in the participant pool of this study allowed us to examine an unresolved issue of critical importance: the role of transfer in the acquisition of anaphoric subjects. Although our methodological setting has been occasionally used in previous literature (Bel, Sagarra, Comínguez, & García-Alcaraz, 2016; Lozano, 2002b, 2002c; Polio, 1995; White, 1985, 1986), this is the first time that different-L1 groups are compared across three (and not one) proficiency levels. This allows significantly safer conclusions to be drawn than in

the case of contrasting a single proficiency level. The major contribution of this study, thus, relies on the fact that the comparison of all three proficiency levels (by L1) points to the same direction: the L1 decisively constrains the L2 acquisition of 3rd person anaphoric subjects. This finding crucially contributes to the ongoing debate on the L2 acquisition of linguistic knowledge at the interfaces, where cross-linguistic influence has often been overlooked in favour of alternative explanations related to processing resources and/or other factors. At the same time, the examination of such an intricate participant pool allowed us to provide some insights regarding the complexity of the phenomenon. We can safely conclude that cross-linguistic influence cannot be the whole story, given the non-target performance of some learners from both L1 backgrounds. Overall, whereas the role of transfer is crucial, our results indicate that processing and/or other factors (input, universal clarity principles, etc.) must also be involved in the acquisition of anaphoric subjects. However, another limitation of the present study relies on our inability to control the particular relevance of each factor for the production of L2 learners. Future research will need to determine the extent to which each factor affects the L2 acquisition of anaphoric subjects. Additionally, our methodological setting allowed us to examine the often overlooked developmental patterns in the L2 acquisition of anaphoric subjects. The present study provided some solid evidence that non-target L2 performance is related to proficiency level, insofar as the more proficient participants of both different-L1 learner groups were found to perform better than their less proficient counterparts. The direct implication of this finding is that any claims made regarding the acquisition of anaphoric subjects should be restricted to the proficiency level of the L2ers under study. In addition, it is of critical importance that the latter is objectively determined through independent proficiency tests. Ideally, in future research, standardized proficiency tests based on the CEFR could be designed and employed to assess the proficiency level of L2 learners. By doing so, we would ensure that the L2 participants of different studies are directly comparable with respect to linguistic competence.

8.2 Limitations and future research

Finally, it is crucial to point out some limitations of the present study that could be overcome in future research:

- First, this study has examined some particular language combinations with a limited number of participants in a specific discourse genre and mode. Our

findings should be tested in the future with other languages combinations as well (e.g. Greek-speaking learners of Italian). Lower proficiency learners (e.g. beginners) should also be included in the participant pool in future studies and more participants per group should be tested. Different discourse genres and modes (e.g. oral conversations) should be examined as well.

- Second, as has been recently pointed out by some authors (Díaz-Negrillo & Thompson, 2013; Lozano, 2009b; Mendikoetxea, 2013), production-oriented research must always be complemented with experimental studies. Recall that the results of the present study concern exclusively the production of anaphoric subjects in discourse. Thus, no claims can be made regarding the interpretation of anaphoric subjects and experimental work in this direction is needed to triangulate corpus findings and provide a full account of anaphora in first and second language acquisition. Based on the corpus-based findings of the present study experiments could be designed in the future in order to test each of the particular properties examined here (e.g. the resolution of anaphoric subjects according to clause type, number of interfering referents, etc.)
- Third, regarding the analysis of the L1 data, the methodological approach adopted here did not allow us to examine neither the interaction of factors with each other nor the potential fine-grained differences between the overt subject forms. Future research on discourse anaphora should benefit from the application of sophisticated statistical models in large corpora, in strict cooperation with experts from other disciplinary fields (e.g. statisticians), in order to provide an account of anaphora in all its complexity.
- Fourth, the role of previously learned foreign languages ('L2 to L3' influence) was not specifically addressed in the present study. Future research should determine the extent to which this factor affects the performance of L2 learners. Additionally, it needs to be determined how the interacting factors that affect the L2 acquisition of anaphoric subjects (as well as the L2 acquisition of any other linguistic phenomenon) are operative for different source/target language pairs and proficiency levels.
- Finally, although not specifically controlled in the present study, individual differences may also be at play in the production of anaphoric subjects (as well as the L2 acquisition of any other linguistic phenomenon). Future qualitative

research should determine the extent to which individual preferences (e.g. style) may affect the referential choices of native speakers and L2 learners.

REFERENCES

Abbott, B. (2010). *Reference*. Oxford: Oxford University Press.

Abreu, L. (2009). *Spanish subject pronoun use by monolinguals, bilinguals and second language learners (Doctoral dissertation)*. University of Florida.

Ädel, A. (2015). Variability in learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 401–423). Cambridge: Cambridge University Press.

Al-Kasey, T., & Pérez-Leroux, A. T. (1998). Second language acquisition of Spanish null subjects. In S. Flynn, G. Martohardjono, & W. A. O'Neil (Eds.), *The generative study of second language acquisition* (pp. 161–185). Hillsdale, NJ: Lawrence Erlbaum.

Almor, A. (2000). Constraints and mechanisms in theories of anaphor processing. In M. Pickering, C. Clifton, & M. Crocker (Eds.), *Architectures and mechanisms for language processing* (pp. 341–354). England: Cambridge University Press.

Alonso-Ovalle, L., Fernández-Solera, S., Frazier, L., & Clifton, C. (2002). Null vs. overt pronouns and the topic-focus articulation in Spanish. *Italian Journal of Linguistics*, *14*(2), 151–170.

Alonso, P. (2006). Discourse strategies for global topic construction in complex written texts : evidence from comment articles. *Revista Alicantina de Estudios Ingleses*, *19*, 9–22.

Andreou, M., Knopp, E., Bongartz, C., & Tsimpli, I. (2015). Character reference in Greek-German bilingual children's narratives. *EUROSLA Yearbook*, *15*(9), 1–40. https://doi.org/10.1075/eurosla.15.01and

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics*, *24*(1), 65–87. https://doi.org/10.1017/S0022226700011567

Ariel, M. (1990). *Accessing noun phrase antecedents*. London: Routledge.

Ariel, M. (1994). Interpreting anaphoric expressions: a cognitive versus a pragmatic approach. *Journal of Linguistics*, *30*(1), 3–42. https://doi.org/10.1017/S0022226700016170

Ariel, M. (1996). Referring expressions and the +/- coreference distinction. In T. Fretheim & J. K. Gundel (Eds.), *Reference and referent accessibility* (pp. 13–35). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Arnold, J. E. (1998). *Reference form and discourse patterns (Doctoral dissertation)*. Stanford University.

Arnold, J. E. (2003). Multiple constraints on reference form. In John W. Du Bois, L. E. Kumpf, & W. J. Ashby (Eds.), *Preferred argument structure: grammar as architecture for function* (pp. 225–245). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Arnold, J. E. (2010). How speakers refer: the role of accessibility. *Language and Linguistics Compass*, *4*(4), 187–203. https://doi.org/10.1111/j.1749-818X.2010.00193.x

Arnold, J. E. (2015). Women and men have different discourse biases for pronoun interpretation. *Discourse Processes*, *52*, 77–110. https://doi.org/10.1080/0163853X.2014.946847

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, *76*(1). https://doi.org/10.1016/S0010-0277(00)00073-1

Arnold, J. E., & Griffin, Z. M. (2007). The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language*, *56*(4), 521–536. https://doi.org/10.1016/j.jml.2006.09.007

Arnold, J. E., Kaiser, E., Kahn, J. M., & Kim, L. K. (2013). Information structure: linguistic, cognitive, and processing approaches. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(4), 403–413. https://doi.org/10.1002/wcs.1234

Asher, N. (1993). *Reference to abstract objects in discourse*. Dordrecht: Kluwer Academic Publishers.

Asher, N., & Vieu, L. (2005). Subordinating and coordinating discourse relations. *Lingua*, *115*, 591–610. https://doi.org/10.1016/j.lingua.2003.09.017

Avila-Jiménez, B. I. (1995). A sociolinguistic analysis of a change in progress: pronominal overtness in Puerto Rican Spanish. *Cornell Working Papers in Linguistics*, *13*, 25–47.

Bardel, C., & Falk, Y. (2007). The role of the second language in third language acquisition: the case of Germanic syntax. *Second Language Research*, *23*(4), 459–484. https://doi.org/10.1177/0267658307080557

Bayley, R., & Langman, J. (2004). Comparing variation in the group and the individual: evidence from second language acquisition. *IRAL*, *42*(4), 303–319.

Bayley, R., & Pease-Alvarez, L. (1997). Null pronoun variation in Mexican-descent children's narrative discourse. *Language Variation and Change*, *9*(3), 349. https://doi.org/10.1017/S0954394500001964

Bayley, R., & Preston, D. R. (1996). *Second language acquisition and linguistic variation*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Beasley, T. M., & Schumacker, R. E. (1995). Multiple regression approach to analyzing contingency tables: post hoc and planned comparison procedures. *The Journal of Experimental Education*, *64*(1), 79–93. https://doi.org/10.1080/00220973.1995.9943797

Beavers, J., & Sag, I. A. (2004). Coordinate ellipsis and apparent non-constituent coordination. In S. Müller (Ed.), *Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar* (pp. 48–69). Stanford: CSLI.

Bel, A., & García-Alcaraz, E. (2015). Subject pronouns in the L2 Spanish of Moroccan Arabic speakers. In T. Judy & S. Perpiñán (Eds.), *The acquisition of Spanish in understudied language pairings* (pp. 201–232). Amsterdam: John Benjamins. https://doi.org/10.1075/ihll.3.08bel

Bel, A., García-Alcaraz, E., & Rosado, E. (2016). Reference comprehension and production in bilingual Spanish: the view from null subject languages. In A. A. de la Fuente, E. Valenzuela, & C. M. Sanz (Eds.), *Language acquisition beyond parameters. Studies in honour of Juana M. Liceras* (pp. 37–70). Amsterdam: John Benjamins.

Bel, A., Perera, J., & Salas, N. (2010). Anaphoric devices in written and spoken narrative discourse: Data from Catalan. *Written Language & Literacy*, *13*(2), 234–259. https://doi.org/10.1075/wll.13.2.03bel

Bel, A., Sagarra, N., Comínguez, J. P., & García-Alcaraz, E. (2016). Transfer and proficiency effects in L2 processing of subject anaphora. *Lingua*, *184*, 134–159. https://doi.org/10.1016/j.lingua.2016.07.001

Belletti, A., Bennati, E., & Sorace, A. (2007). Theoretical and developmental issues in the syntax of subjects: Evidence from near-native Italian. *Natural Language & Linguistic Theory*, *25*(4), 657–689. https://doi.org/10.1007/s11049-007-9026-9

Bentivoglio, P. (1983). Topic continuity and discontinuity in discourse. In T. Givón (Ed.), *Topic continuity in discourse: A quantitative cross-language study* (pp. 255–311). Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.3.06ben

Bentivoglio, P. (1987). *Los sujetos pronominales de primera persona en el habla de Caracas*. Caracas: Universidad Central de Venezuela.

Benveniste, E. (1971). Problems in general Linguistics. In E. Benveniste (Ed.), *The nature of pronouns* (pp. 217–222). Coral Gables FL: University of Miami Press.

Berman, R. (2008). The psycholinguistics of developing text construction. *Journal of Child Language*, *35*(4), 735–771. https://doi.org/10.1017/S0305000908008787

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: investigating language structure and use*. Cambridge: CUP.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson Education limited.

Biberauer, T., Holberg, A., Roberts, I., & Sheehan, M. (2010). *Parametric variation: null subjects in Minimalist Theory*. New York: Cambridge University Press.

Bini, M. (1993). La adquisición del italiano: más allá de las propiedades sintácticas del parámetro pro-drop en el español no nativo. In J. Muñoz-Liceras (Ed.), *La Lingüística y el análisis de los sistemas no nativos* (pp. 126–139). Ottawa: Dovehouse Editions.

Blackwell, S. E. (1998). Constraints on Spanish NP anaphora: the syntactic versus the pragmatic domain. *Hispania*, *81*(3), 606–618. https://doi.org/10.2307/345683

Blackwell, S. E. (2003). *Implicatures in discourse: The case of Spanish NP anaphora*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

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. M. Geeslin & M. Díaz-Campos (Eds.), *Selected Proceedings of the 14th Hispanic Linguistics Symposium* (pp. 142–164). Somerville, MA: Cascadilla Proceedings Project.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, *33*(1), 1–17. https://doi.org/10.1111/j.1467-1770.1983.tb00983.x

Botley, S. (1999). *Corpora and discourse anaphora: using corpus evidence to test theoretical claims (Doctoral dissertation)*. Lancaster University.

Botley, S., & McEnery, T. (2000). *Corpus-based and computational approaches to discourse anaphora*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge: Cambridge University Press.

Büring, D. (2005). *Binding theory*. New York: Cambridge University Press.

Cameron, R. (1994). Switch reference, verb class and priming in a variable syntax. In K. Beals (Ed.), *Papers from the regional meeting of the Chicago Linguistic Society: the parasession on variation in linguistic theory* (pp. 27–45). Chicago: Chicago Linguistic Society.

Cameron, R. (1995). The scope and limits of switch reference as a constraint on pronominal subject expression. *Hispanic Linguistics*, *6/7*, 1–27.

Cameron, R., & Flores-Ferrán, N. (2004). Perseveration of subject expression across regional dialects of Spanish. *Spanish in Context*, *1*(1), 41–65. https://doi.org/10.1075/sic.1.1.05cam

Caramazza, A., Grober, E., Garvey, C., & Yates, J. (1977). Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 601–609. https://doi.org/10.1016/S0022-5371(77)80022-4

Carlsen, C. (2012). Proficiency level: a fuzzy variable in computer Learner Corpora. *Applied Linguistics*. https://doi.org/10.1093/applin/amr047

Carminati, M. N. (2002). *The processing of Italian subject pronouns (Doctoral dissertation)*. University of Massachusetts Amherst.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In W. Chafe (Ed.), *Subject and topic* (pp. 27–55). New York: Academic Press.

Chafe, W. (1980). *The Pear Stories: cognitive, cultural, and linguistic aspects of narrative production*. Norwood, New Jersey: Ablex.

Chafe, W. (1987). Cognitive constraints on information flow. In R. S. Tomlin (Ed.), *Coherence and grounding in discourse* (pp. 21–52). Amsterdam/Philadelphia: John Benjamins Publishing.

Chafe, W. (1990). Introduction to a special issue on third-person reference in discourse. *International Journal of American Linguistics*, *56*(3), 313–316.

Chafe, W. (1994). *Discourse, consciousness and time: the flow and displacement of conscious experience in speaking and writing*. Chicago: The University of Chicago Press.

Charatzidis, A., Georgopoulos, A., Papadopoulou, D., & Tantos, A. (2015). Anaphora resolution in Greek: a corpus-based study. Paper presented at the *12th International Conference on Greek Linguistics (ICGL12)*. Berlin.

Chen, L., & Lei, J. (2012). The production of referring expressions in oral narratives of Chinese-English bilingual speakers and monolingual peers. *Child Language Teaching and Therapy*, *29*(1), 41–55. https://doi.org/10.1177/0265659012459527

Chini, M. (2005). Reference to person in learner discourse. In H. Hendriks (Ed.), *The structure of learner varieties* (pp. 65–110). Berlin: Mouton de Gryter.

Chini, M. (2009). Acquiring the grammar of topicality in L2 Italian : a comparative approach. In L. Mereu (Ed.), *Information structure and its interfaces* (pp. 351–386). Berlin: Mouton de Gryter.

Chomsky, N. (1980). On binding. *Linguistic Inquiry*, *11*(1), 1–46.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin and use*. New York: Praeger.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge: MIT Press.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*, 3–42. https://doi.org/10.1017/S0142716406060024

Clancy, P. M. (1980). Referential choice in English and Japanese narrative discourse. In W. Chafe (Ed.), *The Pear stories: cognitive, cultural and linguistic aspects of narrative production* (pp. 127–202). Norwood, New Jersey: Ablex.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. Joshi, B. Webber, & I. Sagg (Eds.), *Elements of discourse understanding* (pp. 10–

63). Cambridge: Cambridge University Press.

Clark, H. H., & Sengul, C. J. (1979). In search of referents for nouns and pronouns. *Memory & Cognition*, *7*(1), 35–41. https://doi.org/10.3758/BF03196932

Cornish, F. (1999). *Anaphora, discourse and understanding. Evidence from English and French.* Oxford: Oxford University Press.

Cornish, F. (2006). Discourse anaphora. In E. K. Brown & A. Anderson (Eds.), *Encyclopedia of language and linguistics* (2nd ed., pp. 631–638). Amsterdam: Elsevier.

Cornish, F. (2010). Anaphora: text-based or discourse-dependent?: Functionalist vs. formalist accounts. *Functions of Language*, *17*(2), 207–241. https://doi.org/10.1075/fol.17.2.03cor

Council of Europe. (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Press Syndicate of the University of Cambridge.

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, *91*(1), 15–30. https://doi.org/10.1111/j.1540-4781.2007.00507.x

Dahl, Ö., & Fraurud, K. (1996). Animacy in grammar and discourse. In T. Fretheim & J. K. Gundel (Eds.), *Reference and referent accessibility* (pp. 47–64). Amsterdam/Philadelphia: John Benjamins.

De Angelis, G. (2005). Interlanguage transfer of function words. *Language Learning*, *55*(3), 379–414. https://doi.org/10.1111/j.0023-8333.2005.00310.x

Díaz-Negrillo, A., & Fernández Domínguez, J. (2006). Error tagging systems for learner corpora. *Revista Española de Lingüística Aplicada*, *19*, 83–102.

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. 9–30). Amsterdam: John Benjamins.

Dracos, M. (2010). Spanish subject pronoun use in instructional input. Paper presented at the *14th Hispanic Linguistics Symposium*. Bloomington, IN.

Dumont, J. (2006). Full NPs as subjects. In Nuria Sagarra & A. J. Toribio (Eds.), *Selected proceedings of the 9th Hispanic Linguistics Symposium* (pp. 286–296). Somerville, MA: Cascadilla Proceedings Project.

Ellis, R. (1985). *Understanding second language acquisition*. Oxford: Oxford University Press.

Ellis, R. (1989). Sources of intra-learner variability in language use and their relationship to Second Language Acquisition. In S. M. Gass, C. Madden, D. R. Preston, & L. Selinker (Eds.), *Variation in second language acquisition: Vol. II . Psycholinguistic issues* (pp. 22–42). Clevedon, UK: Multilingual Matters.

Emmott, C. (1997). *Narrative comprehension: a discourse perspective*. New York: Oxford University Press.

Emmott, C. (2006). Reference: stylistic aspects. In K. Brown & A. Anderon (Eds.), *Encyclopedia of language and linguistics* (2nd ed., pp. 441–450). Amsterdam: Elsevier.

Eslami Rasekh, A. (1997). *An investigation into discourse anaphoric relations: On the role of contextual information in anaphor resolution (Doctoral dissertation)*. University of Monash.

Fakhri, A. (1989). Variation in the use of referential forms. In S. M. Gass, C. Madden, D. R. Preston, & L. Selinker (Eds.), *Variation in second language acquisition: Vol. II . Psycholinguistic issues* (pp. 189–201). Clevedon, UK: Multilingual Matters.

Falk, Y., & Bardel, C. (2011). Object pronouns in German L3 syntax: evidence for the L2 status factor. *Second Language Research*, *27*(1), 59–82. https://doi.org/10.1177/0267658310386647

Fernandez de Moya, Z. (1996). *La identificación de los sujetos nulos en el español no nativo (Master thesis)*. University of Ottawa.

Fernández Soriano, O. M. (1999). El pronombre personal. Formas y distribuciones. Pronombres átonos y tónicos. In I. Bosque & V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (pp. 1209–1274). Madrid: Espasa Calpe.

Filiaci, F. (2010). *Anaphoric preferences of null and overt subjects in Italian and Spanish: a cross-linguistic comparison (Doctoral dissertation)*. University of Edinburgh. https://doi.org/10.1080/01690965.2013.801502

Filiaci, F., Sorace, A., & Carreiras, M. (2014). Anaphoric biases of null and overt subjects in Italian and Spanish: a cross-linguistic comparison. *Language, Cognition and Neuroscience*, *29*(7), 825–843. https://doi.org/10.1080/01690965.2013.801502

Firbas, J. (1956). “Poznamky k problematice anglickeho slovniho pofadku z hlediska aktualniho cleneni vetneho” (Notes on the problems of English word order from the point of view of functional sentence perspective). In *Proceedings of the philosophical faculty of the University of Brno* (Vol. A4, pp. 93–107). Brno: Masaryk University, Faculty of Arts.

Firbas, J. (1966). On defining the theme in Functional Sentence Analysis. *Travaux Linguistique de Prague*, *1*, 267–280.

Firbas, J. (1992). *Functional sentence perspective in written and spoken communication*. Cambridge: Cambridge University Press.

Flores-Ferrán, N. (2002). *Subject personal pronouns in Spanish narratives of Puerto Ricans in New York City: A sociolinguistic perspective*. Munich: Lincom Europa.

Flores-Ferrán, N. (2004). Spanish subject personal pronoun use in New York City Puerto Ricans: Can we rest the case of English contact? *Language Variation and Change*, *16*(1), 49–73. https://doi.org/10.1017/S0954394504161048

Flores-Ferrán, N. (2010). ¡Tú no me hables! Pronoun expression in conflict narratives. *International Journal of the Sociology of Language*, (203), 61–82. https://doi.org/10.1515/IJSL.2010.022

Fox, B. (1987a). *Discourse structure and anaphora: written and conversational English*. Cambridge: Cambridge University Press.

Fox, B. (1987b). Morpho-syntactic markedness and discourse structure. *Journal of Pragmatics*, *11*(3), 359–375. https://doi.org/10.1016/0378-2166(87)90137-8

Fretheim, T., & Gundel, J. K. (1996). *Reference and referent accessibility*. Amsterdam/Philadelphia: John Benjamins.

Fukumura, K., & van Gompel, R. P. G. (2011). The effect of animacy on the choice of referring expression. *Language and Cognitive Processes*, *26*(10), 1472–1504. https://doi.org/10.1080/01690965.2010.506444

Fukumura, K., & van Gompel, R. P. G. (2015). Effects of order of mention and grammatical role on anaphor resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(2), 501–525. https://doi.org/10.1037/xlm0000041

Gaillat, T. (2016). *Reference in interlanguage: the case of This and That. From linguistic*

annotation to corpus interoperability (Doctoral dissertation). Université Sorbonne Paris.

García-Alcaraz, E. (2015). *Comprensión y producción de los pronombres nulos y explícitos de tercera persona en posición de sujeto en la adquisición temprana del español L2 (Doctoral Dissertation)*. Universidad Pompeu Fabra.

García-Alcaraz, E., & Bel, A. (2011). Selección y distribución de los pronombres en el español L2 de los hablantes de arabe. *Revista de Lingüística Y Lenguas Aplicadas*, *6*, 165–179.

Garcia-Pérez, M. A., & Nuñez-Anton, V. (2003). Cellwise residual analysis in two-way contingency tables. *Educational and Psychological Measurement*, *63*(5), 825–839. https://doi.org/10.1177/0013164403251280

Gardelle, L. (2012). "Anaphora", "anaphor" and "antecedent" in nominal anaphora: definitions and theoretical implications. *Cercles : Revue Pluridisciplinaire Du Monde Anglophone*, (22), 25–40.

Gardelle, L., & Sorlin, S. (2015). *The pragmatics of personal pronouns*. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/slcs.171

Garnham, A., & Cowles, H. W. (2006). Reference : psycholinguistic approach. In E. K. Brown & A. Anderson (Eds.), *Encyclopedia of language and linguistics* (2nd ed., pp. 427–433). Amsterdam: Elsevier.

Garnham, A., & Oakhill, J. (1990). Mental models as contexts for interpreting texts: Implications from studies of anaphora. *Journal of Semantics*, *7*(4), 379–393. https://doi.org/10.1093/jos/7.4.379

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, *5*(3), 459–464.

Gass, S. M., Madden, C., Preston, D. R., & Selinker, L. (1989). *Variation in second language acquisition: Vol. II . Psycholinguistic issues*. Clevedon, UK: Multilingual Matters.

Gass, S. M., & Selinker, L. (1992). *Language transfer in language learning*. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/lald.5

Geeslin, K. L., & Gudmestad, A. (2008). Variable subject expression in second-language Spanish: A comparison of native and non-native speakers. In M. Bowles (Ed.),

Selected proceedings of the 2007 second language research forum (pp. 69–85). Somerville: MA: Cascadilla Proceedings Project.

Geeslin, K. L., & Gudmestad, A. (2011). Using sociolinguistic analyses of discourse-level features to expand research on L2 variation in forms of Spanish subject expression. In L. Plonsky & M. Schierloh (Eds.), *Selected proceedings of the 2009 second language research forum* (pp. 16–30). Somerville, MA: Cascadilla Proceedings Project.

Geeslin, K. L., & Gudmestad, A. (2016). Subject expression in Spanish: contrasts between native and non-native speakers for first and second-person singular referents. *Spanish in Context*, *13*(1), 53–79. https://doi.org/10.1075/sic.13.1.03gee

Geeslin, K. L., Linford, B., & Fafulas, S. (2015). Variable subject expression in second language Spanish. In A. M. Carvalho, R. Orozco, & N. L. Shin (Eds.), *Subject pronoun expression in Spanish: a cross-dialectal perspective* (pp. 191–209). Washington, D.C.: Georgetown University Press.

Geluykens, R. (2013). *Pragmatics of discourse anaphora in English: evidence from conversational repair*. Tübingen: Walter de Gruyter.

Gernsbacher, M. A. (1989). Mechanisms that improve referential access. *Cognition*, *32*(2), 99–156. https://doi.org/10.1016/0010-0277(89)90001-2

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gibson, E., & Pearlmutter, N. J. (2011). *The processing and acquisition of reference*. Cambridge, Massachusetts: MIT Press.

Givón, T. (1983). *Topic continuity in discourse*. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/tsl.3

Givón, T. (1990). *Syntax: a functional-typological introduction*. Amsterdam: John Benjamins.

Goikoetxea, E., Pascual, G., & Acha, J. (2008). Normative study of the implicit causality of 100 interpersonal verbs in Spanish. *Behavior Research Methods*, *40*(3), 760–772. https://doi.org/10.3758/BRM.40.3.760

Granger, S. (2002). A bird's-eye view of learner corpus research. In J. Hung, S. Petch-Tyson, & S. Granger (Eds.), *Computer learner corpora, second language*

acquisition and foreign language teaching (pp. 3–33). Amsterdam: John Benjamins Publishing. https://doi.org/10.2307/1315064

Granger, S. (2004). Computer learner corpus research: current status and future prospects. *Language and Computers*, *52*(1), 123–145.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13–23). Amsterdam: John Benjamins.

Granger, S. (2012). How to use foreign and second language learner corpora. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: a practical guide* (pp. 7–29). London: Blackwell Publishing.

Granger, S., Gilquin, G., & Meunier, F. (2015). *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Granger, S., Hung, J., & Petch-Tyson, S. (2002). *Computer learner corpora, Second Language Acquisition and Foreign Language Learning*. Amsterdam: John Benjamins.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics 3: Speech arts* (pp. 41–58). New York: Academic Press.

Gries, S. T. (2012). Corpus Linguistics: quantitative methods. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1380–1385). Oxford: Wiley-Blackwell.

Gries, S. T. (2015). Statistics for learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 159–183). Cambridge: Cambridge University Press.

Gries, S. T., & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora*, *9*(1), 109–136. https://doi.org/10.3366/cor.2014.0053

Grimes, J. E. (1978). *Papers in discourse*. Arlington: SIL.

Grüning, A., & Kibrik, A. A. (2005). Modelling referential choice in discourse: a cognitive calculative approach and a neural networks approach. In A. Branco, T. McEnery, & R. Mitkov (Eds.), *Anaphora processing: linguistic, cognitive and computational modelling* (pp. 163–198). Amsterdam: John Benjamins.

Gudmestad, A., & Geeslin, K. L. (2010). Exploring the roles of redundancy and ambiguity in variable subject expression: A comparison of native and non-native speakers. In C. Borgonovo (Ed.), *Selected proceedings of the 12th Hispanic linguistics symposium* (pp. 270–283). Somerville MA: Cascadilla Proceedings Project.

Gudmestad, A., House, L., & Geeslin, K. L. (2013). What a Bayesian analysis can do for SLA: new tools for the sociolinguistic study of subject expression in L2 Spanish. *Language Learning*, *63*(3), 371–399. https://doi.org/10.1111/lang.12006

Gundel, J. K. (1974). *The role of topic and comment in linguistic theory (Doctoral dissertation)*. University of Texas.

Gundel, J. K. (2010). Reference and accessibility from a Givenness Hierarchy perspective. *International Review of Pragmatics*, *2*(2), 148–168. https://doi.org/10.1163/187731010X528322

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*(2), 274–307. https://doi.org/10.2307/416535

Gundel, J. K., Stenson, N., & Tarone, E. (1984). Acquiring pronouns in a second language: evidence for hypothesis testing. *Studies in Second Language Acquisition*, *6*(2), 215–225. https://doi.org/10.1017/S0272263100005027

Gundel, J. K., & Tarone, E. (1992). "'Language transfer'" and the acquisition of pronominal anaphora. In S. M. Gass & L. Selinker (Eds.), *Language transfer in language learning* (pp. 87–101). Amsterdam: John Benjamins Publishing Company.

Haegeman, L. (1991). *Introduction to Government and Binding Theory*. Oxford: Blackwell.

Haegeman, L. (1997). Register variation, truncation, and subject omission in English and in French. *English Language and Linguistics*, *1*(2). https://doi.org/10.1017/S1360674300000526

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hartshorne, J. K., Sudo, Y., & Uruwashi, M. (2013). Are implicit causality pronoun resolution biases consistent across languages and cultures? *Experimental Psychology*, *60*(3), 179–196. https://doi.org/10.1027/1618-3169/a000187

Hendriks, H. (2003). Using nouns for reference maintenance: A seeming contradiction in L2 discourse. In A. G. Ramat (Ed.), *Typology and second language acquisition* (pp. 291–326). Berlin: Walter de Gruyter.

Hendriks, P., Koster, C., & Hoeks, J. C. J. (2014). Referential choice across the lifespan: why children and elderly adults produce ambiguous pronouns. *Language, Cognition and Neuroscience*, *29*(4), 391–407. https://doi.org/10.1080/01690965.2013.766356

Hinds, J. (1977). Paragraph structure and pronominalization. *Paper in Linguistics*, *10*(1–2), 77–99. https://doi.org/10.1080/08351819709370440

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, *3*(1), 67–90. https://doi.org/10.1207/s15516709cog0301_4

Hockett, C. F. (1958). *A course in modern linguistics*. New York: Macmillan.

Hofmann, T. R. (1989). Paragraphs and anaphora. *Journal of Pragmatics*, *13*(2), 239–250. https://doi.org/10.1016/0378-2166(89)90093-3

Huang, Y. (2000a). *Anaphora: a cross-linguistic approach*. New York: Oxford University Press.

Huang, Y. (2000b). Discourse anaphora: four theoretical models. *Journal of Pragmatics*, *32*(2), 151–176. https://doi.org/10.1016/S0378-2166(99)00041-7

Hurtado, L. M. (2005). Condicionamientos sintáctico-semánticos de la expresión del sujeto en el español colombiano. *Hispania*, *88*(2), 335–348.

Ivanov, I. P. (2012). L2 acquisition of Bulgarian clitic doubling: A test case for the Interface Hypothesis. *Second Language Research*, *28*(3), 345–368. https://doi.org/10.1177/0267658312452066

Jackendoff, R. (1997). *The architecture of the language faculty*. Cambridge, Massachusetts: MIT Press.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Cambridge, Massachusetts: MIT Press.

Jaeggli, O. (1982). *Topics in Romance Syntax*. Dordrecht: Foris.

Jaeggli, O., & Safir, K. (1989). The null subject parameter and parametric theory. In O. Jaeggli & K. Safir (Eds.), *The Null Subject Parameter* (pp. 1–44). Dordrecht, Boston, London: Kluwer Academic Publishers. https://doi.org/10.1007/978-94-009-2540-3_1

Jegerski, J., VanPatten, B., & Keating, G. D. (2011). Cross-linguistic variation and the acquisition of pronominal reference in L2 Spanish. *Second Language Research*, *27*(4), 481–507. https://doi.org/10.1177/0267658311406033

Jespersen, O. (1924). *The philisophy of grammar*. London: Allen and Unwin.

Judy, T. (2015). Knowledge and processing of subject-related discourse properties in L2 near-native speakers of Spanish, L1 Farsi. In T. Judy & Silvia Perpiñán (Eds.), *The acquisition of Spanish in understudied language pairings* (pp. 169–200). Amsterdam: John Benjamins. https://doi.org/10.1075/ihll.3.07jud

Kaiser, E. (2003). *The quest for a referent: A crosslinguistic look at reference resolution (Doctoral dissertation)*. University of Pennsylvania.

Kaiser, E., Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2009). Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition*, *112*(1), 55–80. https://doi.org/10.1016/j.cognition.2009.03.010

Kaiser, E., & Trueswell, J. C. (2008). Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes*, *23*(5), 709–748. https://doi.org/10.1080/01690960701771220

Kang, J. Y. (2004). Telling a coherent story in a foreign language: analysis of Korean EFL learners' referential strategies in oral narrative discourse. *Journal of Pragmatics*, *36*(11), 1975–1990. https://doi.org/10.1016/j.pragma.2004.03.007

Keating, G. D., VanPatten, B., & Jegerski, J. (2011). Who was walking on the beach? *Studies in Second Language Acquisition*, *33*(2), 193–221. https://doi.org/10.1017/S0272263110000732

Kennedy, C., & Bolitho, R. (1984). *English for specific purposes*. Hong Kong: Macmillan.

Kibrik, A. A. (1996). Anaphora in Russian narrative prose: a cognitive calculative account. In B. Fox (Ed.), *Studies in anaphora* (pp. 255–305). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Kibrik, A. A. (2000). A cognitive calculative approach towards discourse anaphora. In P. Baker, A. Hardie, T. McEnery, & A. Siewierska (Eds.), *Proceedings of the discourse anaphora and reference resolution conference (DAARC 2000)* (pp. 72–82). Lancaster University: University Centre for Computer Corpus Research on

Language.

Kibrik, A. A. (2001). Reference maintenance in discourse. In M. Haspelmath (Ed.), *Language typology and language universals* (Vol. 2, pp. 1123–1141). Berlin: Mouton de Gryter.

Kibrik, A. A. (2011). *Reference in discourse*. New York: Oxford University Press.

Kibrik, A. A., Khudyakova, M. V., Dobrov, G. B., Linnik, A., & Zalmanov, D. A. (2016). Referential choice: predictability and its limits. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01429

Kleiber, G. (1994). *Anaphores et pronoms*. Louvain-la-Neuve: Duculot.

Kras, T. (2008). Anaphora resolution in near-native Italian grammars: evidence from native speakers of Croatian. *Eurosla Yearbook*, *8*(1), 107–134.

Kuno, S. (1972). Functional sentence perspective: a case study from Japanese and English. *Linguistic Inquiry*, *3*(3), 269–320. https://doi.org/10.2307/4177715

Labov, W. (1966). *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula. *Language*, *45*(4), 715–762. https://doi.org/10.2307/412333

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University de Pennsylvania Press.

Lambrecht, K. (1994). *Information structure and sentence form. Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Leal Mendez, T., & Slabakova, R. (2014). The Interpretability Hypothesis again: A partial replication of Tsimpli and Dimitrakopoulou (2007). *International Journal of Bilingualism*, *18*(6), 537–557. https://doi.org/10.1177/1367006912448125

Leclercq, P., & Lenart, E. (2013). Discourse cohesion and accessibility of referents in oral narratives: a comparison of L1 and L2 acquisition of French and English. *Discours*, *12*, 3–31.

Levinson, S. C. (1991). Pragmatic reduction of the Binding conditions revisited. *Journal of Linguistics*, *27*(1), 107–161. https://doi.org/10.1017/S0022226700012433

Levinson, S. C. (1995). Three levels of meaning. In F. Palmer (Ed.), *Grammar and*

meaning (pp. 90–115). Cambridge: Cambridge University Press.

Liceras, J. M. (1988). Syntax and stylistics: more on the pro-drop parameter. In J. Pankhurst, M. S. Smith, & P. Van Buren (Eds.), *Learnability and second languages: A book of readings* (pp. 71–93). Dordrecht: Foris.

Liceras, J. M. (1989). On some properties of the pro-drop parameter: looking for missing subjects in non-native Spanish. In S. Gass & J. Schachter (Eds.), *Lingusitic perspectives on second language acquisition* (pp. 109–133). Cambridge: CUP.

Liceras, J. M., de la Fuente, A. A., & Sanz, C. M. (2010). The distribution of null subjects in non-native grammars: syntactic markedness and interface vulnerability. In M. Iverson, I. Ivanov, T. Judy, J. Rothman, R. Slabakova, & M. Tryzna (Eds.), *Proceedings of the 2009 Mind/Context Divide Workshop* (pp. 84–95). Somerville MA: Cascadilla Proceedings Project.

Liceras, J. M., & Díaz, L. (1999). Topic-drop versus pro-drop: null subjects and pronominal subjects in the Spanish L2 of Chinese, English, French, German and Japanese speakers. *Second Language Research*, *15*, 1–40.

Linford, B., & Shin, N. L. (2013). Lexical frequency effects on L2 Spanish subject pronoun expression. In J. C. Amaro (Ed.), *Selected Proceedings of the 16th Hispanic linguistics symposium* (pp. 175–189). Somerville, MA: Cascadilla Proceedings Project.

Long, M. H., & Porter, P. a. (1985). Group work, interlanguage talk, and second language acquisition. *TESOL Quarterly*, *19*(2), 207–228. https://doi.org/10.2307/3586827

Loukachevitch, Natalia V. Khudyakova, M. V., Kibrik, A. A., Dobrov, G. B., & Linnik, A. S. (2011). Computational modeling of referential choice: major and minor referential options. In *Proceedings of the PRE-CogSci 2011 workshop on production of referring expressions: bridging the gap between computational, empirical & theoretical approaches* (pp. 458–468). Boston.

Lozano, C. (2002a). Knowledge of expletive and pronominal subjects by learners of Spanish. *ITL Review of Applied Linguistics*, *135*(6), 37–60.

Lozano, C. (2002b). Pronominal mental representations in advanced L2 and L3 learners of Spanish. In J. D. Luque Durán, A. Pamiés Bertrán, & F. Manjón Pozas (Eds.), *Nuevas tendencias en la investigación lingüística* (pp. 605–617). Granada: Granada Lingüística.

Lozano, C. (2002c). The interpretation of overt and null pronouns in non-native Spanish. *Durham Working Papers in Linguistics*, *8*, 53–66.

Lozano, C. (2008a). ¿Deficits de representación o de procesamiento en una segunda lengua? Evidencia de un estudio de resolución de anáfora con griegos adultos aprendices de español. In R. Monroy & A. Sánchez (Eds.), *25 años de Linguistica Aplicada en España: Hitos y retos* (pp. 855–866). Murcia: Editum.

Lozano, C. (2008b). *The acquisition of syntax and discourse: pronominals and word order in English and Greek learners of Spanish*. Saarbrücken: VDM Verlag.

Lozano, C. (2009a). CEDEL2: Corpus Escrito del Español L2. In B. Callejas & C. M. (Eds.), *Applied linguistics now understanding language and mind /La lingüística aplicada actual : comprendiendo el lenguaje y la mente* (pp. 197–212). Almeria: Universidad de Almer.a.

Lozano, C. (2009b). Selective deficits at the syntax-discourse interface: Evidence from the CEDEL2 corpus. In N. Snape, Y.-K. I. Leung, & M. S. Smith (Eds.), *Representational deficits in SLA : studies in honor of Roger Hawkins* (pp. 127–166). Amsterdam and New York: John Benjamins Publishing.

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: state of the art and perspectives* (pp. 236–275). Amsterdam: John Benjamins. https://doi.org/10.1017/CBO9781107415324.004

Lozano, C. (forthcoming). The development of anaphora resolution at the syntax-discourse interface : pronominal subjects in Greek learners of Spanish. *Journal of Psycholinguistic Research* (Special Issue: Reference and Anaphora in Iberian Languages).

Lozano, C., & Díaz-Negrillo, A. (submitted). Using learner corpus methods in L2 acquisition research : the morpheme order studies. *RESLA*.

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). Amsterdam: John Benjamins.

Luján, M. (1999). Expresión y omisión del pronombre personal. In I. Bosque & V.

Demonte (Eds.), *Gramática descriptiva de la lengua española* (pp. 1275–1316). Madrid: Espasa Calpe.

Lyons, J. (1968). *Introduction to theoretical linguistics*. Cambridge: Cambridge University Press.

Margaza, P., & Bel, A. (2006). Null subjects at the syntax–pragmatics interface: evidence from Spanish interlanguage of Greek speakers. In M. G. O'Brien, C. Shea, & J. Archibald (Eds.), *Proceedings of the 8th generative approaches to Second Language Acquisition conference (GASLA 2006)* (pp. 88–97). Somerville MA: Cascadilla Proceedings Project.

Mathesius, V. (1975). *A functional analysis of present day English*. The Hague: Mouton.

McDonald, J. H. (2014). *Handbook of biological statistics* (3rd ed.). Baltimore, Maryland: Sparky House Publishing.

McKoon, G., Greene, S. B., & Ratcliff, R. (1993). Discourse models, pronoun resolution, and the implicit causality of verbs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1040–1052. https://doi.org/10.1037/0278-7393.19.5.1040

Mendikoetxea, A. (2013). Corpus-based research in second language Spanish. In K. L. Geeslin (Ed.), *The handbook of Spanish Second Language Acquisition* (pp. 11–29). Malden, MA: Wiley-Blackwell.

Miltsakaki, E. (2002). Toward an aposynthesis of topic continuity and intrasentential anaphora. *Computational Linguistics*, *28*(3), 319–355. https://doi.org/10.1162/089120102760276009

Miltsakaki, E. (2007). A rethink of the relationship between salience and anaphora resolution. In A. Branco, T. McEnery, R. Mitkov, & F. Silva (Eds.), *Proceedings of the 6th discourse anaphora and anaphor resolution colloquium (DAARC 2007)* (pp. 91–96). Porto: Centro de Linguística de Universidade do Porto.

Mitkov, R. (1994). An integrated model for anaphora resolution. In *Proceedings of the 15th conference on Computational linguistics* (Vol. 2, pp. 1170–1177). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/991250.991342

Mitkov, R. (2002). *Anaphora resolution*. London: Longman.

Montalbetti, M. (1984). *After Binding (Doctoral dissertation)*. MIT.

Montrul, S. (2004a). Subject and object expression in Spanish heritage speakers: A case of morphosyntactic convergence. *Bilingualism: Language and Cognition*, *7*(2), 125–142. https://doi.org/10.1017/S1366728904001464

Montrul, S. (2004b). *The acquisition of Spanish. Morphosyntactic development in monolingual and bilingual L1 acquisition and in adult L2 acquisition*. Amsterdam: John Benjamins.

Montrul, S. (2011). Multiple interfaces and incomplete acquisition. *Lingua*, *121*(4), 591–604. https://doi.org/10.1016/j.lingua.2010.05.006

Montrul, S., & Rodríguez Louro, C. (2006). Beyond the syntax of the null subject parameter. A look at the discourse-pragmatic distribution of null and overt subjects by L2 learners of Spanish. In V. Torrens & L. Escobar (Eds.), *Acquisition of syntax in Romance languages* (Vol. 41, pp. 413–430). Amsterdam: John Benjamins.

Muñoz, C. (2001). La sobre-explicitación de la referencia personal en las narrativas en segunda lengua. In A. I. M. Fernández & V. Colwell (Eds.), *Perspectivas recientes sobre el discurso* (pp. 159–180). León: Universidad de León.

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). Amsterdam and New York: John Benjamins Publishing.

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. 309–330). Cambridge: Cambridge University Press.

Nariyama, S. (2004). Subject ellipsis in English. *Journal of Pragmatics*, *36*(2), 237–264. https://doi.org/10.1016/S0378-2166(03)00099-7

Nicol, J. L., & Swinney, D. A. (2003). The psycholinguistics of anaphora. In A. Barss (Ed.), *Anaphora: A reference guide* (pp. 72–104). Malden, MA, USA: Blackwell Publishing Ltd. https://doi.org/10.1002/9780470755594

O'Donnell, M. (2009). The UAM CorpusTool: software for corpus annotation and exploration. In Carmen María Bretones Callejas, J. F. F. Sánchez, J. R. I. Ibáñez, M.

E. G. Sánchez, M. E. C. de los Ríos, M. S. S. Ramiro, … B. C. Márquez (Eds.), *Applied Linguistics now: understanding language and mind/La Lingüística Aplicada actual: comprendiendo el lenguaje y la mente* (pp. 1433–1447). Almeria: Universidad de Almeria.

Otheguy, R., & Zentella, A. C. (2012). *Spanish in New York. The multilingual apple: Languages in New York City*. New York: Oxford University Press.

Otheguy, R., Zentella, A. C., & Livert, D. (2007). Language and dialect contact in Spanish in New York: Toward the formation of a speech community. *Language*, *83*(4), 770–802. https://doi.org/10.1353/lan.2008.0019

Papadopoulou, D., Peristeri, E., Plemenou, E., Marinis, T., & Tsimpli, I. (2015). Pronoun ambiguity resolution in Greek: Evidence from monolingual adults and children. *Lingua*, *155*, 98–120. https://doi.org/10.1016/j.lingua.2014.09.006

Paradis, J., & Navarro, S. (2003). Subject realization and crosslinguistic interference in the bilingual acquisition of Spanish and English: what is the role of the input? *Journal of Child Language*, *30*(2), 371–393. https://doi.org/10.1017/S0305000903005609

Perales, S., & Portillo, R. (2007). Sobre las propiedades referenciales de los sujetos nulos y pronominales del español oral y escrito. In *Las destrezas orales en la enseñanza del español L2-LE: XVII Congreso Internacional de la Asociación del Español como lengua extranjera (ASELE)* (pp. 889–900). Logroño: Universidad de La Rioja.

Pérez-Leroux, A. T., & Glass, W. R. (1997). OPC effects on the L2 acquisition of Spanish. In A. T. Pérez-Leroux & W. R. Glass (Eds.), *Contemporary perspectives on the acquisition of Spanish* (pp. 149–165). Somerville MA: Cascadilla Press.

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.

Perlmutter, D. (1971). *Deep and surface structure constraints in syntax*. New York: Holt, Rinehart and Winston.

Phinney, M. (1987). The pro-drop parameter in second language acquisition. In T. Roeper & E. Williams (Eds.), *Parameter setting* (pp. 221–238). Dordrecht: D. Reidel Publishing Company. https://doi.org/10.1007/978-94-009-3727-7_10

Pladevall-Ballester, E. (2009). Child L2 development of syntactic and discourse

properties of Spanish subjects. *Bilingualism: Language and Cognition*, *13*(2), 185–216. https://doi.org/10.1017/S1366728909990447

Pladevall-Ballester, E. (2013). Adult L2 Spanish development of syntactic and discourse subject properties in an instructional setting. *RAEL: Revista Electrónica de Lingüística Aplicada*, *12*, 111–129.

Polio, C. (1995). Acquiring Nothing? The use of zero pronouns by nonnative speakers of Chinese and the implications for the acquisition of nominal reference. *Studies in Second Language Acquisition*, *17*(3), 353–377. https://doi.org/10.1017/S0272263100014248

Prentza, A. (2010). *Feature interpretability in Second Language Acquisition: evidence from the Null Subject Parameter in the Greek / English interlanguage (Doctoral dissertation)*. Aristotle University of Thessaloniki.

Prentza, A. (2014a). Can Greek learners acquire the overt subject property of English? A pilot study. *Theory and Practice in Language Studies*, *4*(9), 1770–1777. https://doi.org/10.4304/tpls.4.9.1770-1777

Prentza, A. (2014b). Pronominal subjects in English L2 acquisition and in L1 Greek : issues of interpretation, use and L1 transfer. In Nikolaos Lavidas, T. Alexiou, & A. M. Sougari (Eds.), *Major trends in theoretical and applied Linguistics 2: selected papers from the 20th ISTAL* (pp. 369–386). London: Walter de Gruyter.

Prince, E. F. (1981a). Topicalization, focus-movement, and Yiddish-movement: a pragmatic differentiation. *Berkeley Linguistics Society*, *7*, 249–264. https://doi.org/10.3765/bls.v7i0.2092

Prince, E. F. (1981b). Towards a taxonomy of given/new information. In P. Cole (Ed.), *Radical pragmatics* (pp. 223–254). New York: Academic Press.

Quesada, M. L. (2015). *The L2 acquisition of Spanish subjects*. Boston, Berlin: Walter de Gruyter. https://doi.org/10.1515/9781614514367

Quesada, M. L., & Blackwell, S. E. (2009). The L2 acquisition of null and overt Spanish subject pronouns: a pragmatic approach. In J. Collentine (Ed.), *Selected proceedings of the 11th Hispanic Linguistics symposium* (pp. 117–130). Somerville MA: Cascadilla Proceedings Project.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London and New York: Longman.

https://doi.org/10.2307/415437

Ramchand, G., & Reiss, C. (2007). *The Oxford handbook of linguistic interfaces*. Oxford: Oxford University Press.

Rankin, T. (2015). Learner corpora and grammar. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 231–252). Cambridge: Cambridge University Press.

Regan, V. (2004). The relationship between the group and the individual and the acquisition of native speaker variation patterns: A preliminary study. *IRAL*, *42*(4), 335–348.

Reinhart, T. (1981). Pragmatics and linguistics: an analysis of sentence topics. *Philosophica*, *27*(1), 53–94.

Reinhart, T. (1983). Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy*, *6*(1), 47–88. https://doi.org/10.1007/BF00868090

Rizzi, L. (1986). Null objects in Italian and the theory of pro. *Linguistic Inquiry*, *17*(3), 501–557. https://doi.org/10.2307/4178501

Rohde, H., & Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, *29*(8), 912–927. https://doi.org/10.1080/01690965.2013.854918

Rosengren, P. (1974). *Presencia y ausencia de los pronombres personales sujetos en español moderno*. Stockholm: Acta Universitatis Gothoburgensis.

Rothman, J. (2007). Pragmatic solutions for syntactic problems: understanding some L2 syntactic errors in terms of discourse-pragmatic deficit. In S. Baauw, F. Dirjkoningen, & M. Pinto (Eds.), *Romance languages and linguistic theory* (pp. 299–320). Amsterdam: John Benjamins.

Rothman, J. (2009). Pragmatic deficits with syntactic consequences?: L2 pronominal subjects and the syntax–pragmatics interface. *Journal of Pragmatics*, *41*(5), 951–973. https://doi.org/10.1016/j.pragma.2008.07.007

Rothman, J., & Iverson, M. (2007a). Input type and parameter resetting: Is naturalistic input necessary? *IRAL - International Review of Applied Linguistics in Language Teaching*, *45*(4), 285–319. https://doi.org/10.1515/IRAL.2007.013

Rothman, J., & Iverson, M. (2007b). On parameter clustering and resetting the null-subject parameter in L2 Spanish: Implications and observations. *Hispania*, *90*(2), 328–341.

Rothman, J., & Iverson, M. (2007c). The Syntax of null subjects in L2 Spanish: comparing two L2 populations under different exposure. *Revista Española de Lingüística Aplicada*, *20*, 185–214.

Rothman, J., & Slabakova, R. (2011). The mind-context divide: On acquisition at the linguistic interfaces. *Lingua*, *121*(4), 568–576. https://doi.org/10.1016/j.lingua.2011.01.003

Ryan, J. (2015). Overexplicit referent tracking in L2 English: strategy, avoidance, or myth? *Language Learning*, *65*(4), 824–859. https://doi.org/10.1111/lang.12139

Sanford, A. J., & Garrod, S. C. (1981). *Understanding written language*. Chichester: John Wiley and Sons.

Saunders, J. K. (1999). *Null and overt references in Spanish second language acquisition: A discourse perspective (Doctoral dissertation)*. University of Texas.

Schmolz, H. (2015). *Anaphora resolution and text retrieval. A linguistic analysis of hypertexts*. Berlin/München/Boston: De Gruyter Mouton.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, *5*(1), 209–231.

Serratrice, L. (2007a). Cross-linguistic influence in the interpretation of anaphoric and cataphoric pronouns in English–Italian bilingual children. *Bilingualism: Language and Cognition*, *10*(3), 225–238. https://doi.org/10.1017/S1366728907003045

Serratrice, L. (2007b). Referential cohesion in the narratives of bilingual English-Italian children and monolingual peers. *Journal of Pragmatics*, *39*(6), 1058–1087. https://doi.org/10.1016/j.pragma.2006.10.001

Serratrice, L., Sorace, A., & Paoli, S. (2004). Crosslinguistic influence at the syntax–pragmatics interface: Subjects and objects in English–Italian bilingual and monolingual acquisition. *Bilingualism: Language and Cognition*, *7*(3), 183–205. https://doi.org/10.1017/S1366728904001610

Shin, N. L. (2012). Variable use of Spanish subject pronouns by monolingual children in Mexico. In K. L. Geeslin & M. Díaz-Campos (Eds.), *Selected Proceedings of the*

14th Hispanic Linguistics Symposium (pp. 130–141). Somerville MA: Cascadilla Proceedings Project.

Shin, N. L., & Cairns, H. S. (2009). Subject Pronouns in Child Spanish and Continuity of Reference. In J. Collentine (Ed.), *Selected Proceedings of the 11th Hispanic Linguistics Symposium* (pp. 155–164). Somerville MA: Cascadilla Proceedings Project.

Shin, N. L., & Cairns, H. S. (2012). The development of NP selection in school-age children: reference and Spanish subject pronouns. *Language Acquisition*, *19*(1), 3–38. https://doi.org/10.1080/10489223.2012.633846

Shin, N. L., & Erker, D. (2015). The emergence of structured variability in morphosyntax. In A. M. Carvalho, R. Orozco, & N. L. Shin (Eds.), *Subject pronoun expression in Spanish* (pp. 171–193). Washington, D.C.: Georgetown University Press.

Shin, N. L., & Montes-Alcalá, C. (2014). El uso contextual del pronombre sujeto como factor predictivo de la influencia del inglés en el español de Nueva York [English influence on Spanish in New York: Evidence from subject pronouns in context]. *Sociolinguistic Studies*, *8*(1), 85–110. https://doi.org/10.1558/sols.v8i1.85

Shin, N. L., & Otheguy, R. (2009). Shifting sensitivity to Continuity of reference: Subject pronoun use in Spanish in New York City. In M. Lacorte & J. Leeman (Eds.), *Español en Estados Unidos y en otros contextos: Cuestiones sociolingüísticas, políticas y pedagógicas* (pp. 111–136). Madrid: Iberoamericana.

Sidner, C. L. (1981). Focusing for interpretation of pronouns. *American Journal of Computational Linguistics*, *7*(4), 217–232.

Sidner, C. L. (1983). Focusing and discourse. *Discourse Processes*, *6*(2), 107–130. https://doi.org/10.1080/01638538309544558

Silva, V. L. P. (1993). Subject omission and functional compensation: Evidence from written Brazilian Portuguese. *Language Variation and Change*, *5*(1), 35–49. https://doi.org/10.1017/S0954394500001381

Silva Corvalán, C. (1982). Subject expression and placement in Mexican-American Spanish. In J. Amastae & L. Elias-Olivares (Eds.), *Spanish in the United States: sociolinguistic aspects* (pp. 93–120). Cambridge: Cambridge University Press.

Silva Corvalán, C. (1994). *Language contact and change: Spanish in Los Angeles*. Oxford: Clarendon.

Simpson, J., & Weiner, E. (1989). *The Oxford English dictionary*. Oxford: Clarendon Press.

Sinclair, J. (2005). Corpus and text-basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: a guide to good practice* (pp. 1–16). Oxford: Oxbow Books.

Slabakova, R., & García Mayo, M. D. P. (2015). The L3 syntax–discourse interface. *Bilingualism: Language and Cognition*, *18*(2), 208–226. https://doi.org/10.1017/S1366728913000369

Slabakova, R., Kempchinsky, P., & Rothman, J. (2012). Clitic-doubled left dislocation and focus fronting in L2 Spanish: A case of successful acquisition at the syntax-discourse interface. *Second Language Research*, *28*(3), 319–343. https://doi.org/10.1177/0267658312447612

Sorace, A. (2000). Syntactic optionality in non-native grammars. *Second Language Research*, *16*(2), 93–102. https://doi.org/10.1191/026765800670666032

Sorace, A. (2004). Native language attrition and developmental instability at the syntax-discourse interface: Data, interpretations and methods. *Bilingualism: Language and Cognition*, *7*(2), 143–145. https://doi.org/10.1017/S1366728904001543

Sorace, A. (2005). Selective optionality in language development. In L. M. E. A. Cornips & K. P. Corrigan (Eds.), *Syntax and variation: reconciling the biological and the social* (pp. 55–80). Amsterdam/Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/cilt.265.04sor

Sorace, A. (2006a). Possible manifestations of shallow processing in advanced second language speakers. *Applied Psycholinguistics*, *27*, 43–105. https://doi.org/10.1017/S0142716406340039

Sorace, A. (2006b). The use of Acceptability Judgments in second language acquisition research. In W. C. Richie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 375–413). San Diego, California: Academic Press.

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism*, *1*(1), 1–33. https://doi.org/10.1075/lab.2.2.04sor

Sorace, A. (2012). Pinning down the concept of interface in bilingual development: A reply to peer commentaries. *Linguistic Approaches to Bilingualism*, *2*(2), 209–216. https://doi.org/10.1075/lab.2.2.04sor

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

Sorace, A., & Filiaci, F. (2006). Anaphora resolution in near-native speakers of Italian. *Second Language Research*, *22*(3), 339–368. https://doi.org/10.1191/0267658306sr271oa

Sorace, A., & Serratrice, L. (2009). Internal and external interfaces in bilingual language development: Beyond structural overlap. *International Journal of Bilingualism*, *13*(2), 195–210. https://doi.org/10.1177/1367006909339810

Sorace, A., Serratrice, L., Filiaci, F., & Baldo, M. (2009). Discourse conditions on subject pronoun realization: Testing the linguistic intuitions of older bilingual children. *Lingua*, *119*(3), 460–477. https://doi.org/10.1016/j.lingua.2008.09.008

Sperber, D., & Wilson, D. (1986). *Relevance: communication and cognition*. Oxford: Blackwell.

Stevenson, R., Crawley, R., & Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, *9*(4), 519–548. https://doi.org/10.1080/01690969408402130

Stevenson, R., Knott, A., Oberlander, J., & McDonald, S. (2000). Interpreting pronouns and connectives: Interactions among focusing, thematic roles and coherence relations. *Language and Cognitive Processes*, *15*(3), 225–262. https://doi.org/10.1080/016909600386048

Stirling, L., & Huddleston, R. (2010). Deixis and anaphora. In R. Huddleston & G. Pullum (Eds.), *The Cambridge grammar of the English language* (pp. 1449–1564). Cambridge: Cambridge University Press.

Stockwell, R. P., Bowen, J. D., & Martin, J. W. (1965). *The grammatical structures of English and Spanish*. Chicago: University of Chicago Press.

Sullivan, A. (2006). Reference: philosophical theories. In K. Brown & A. Anderson (Eds.), *Encyclopedia of language and linguistics* (2nd ed., pp. 420–427). Amsterdam: Elsevier. https://doi.org/10.1016/B0-08-044854-2/01138-X

Sun, C.-F., & Givón, T. (1985). On the so-called Sov word order in Mandarin Chinese: a quantified text study and its implications. *Language*, *61*(2), 329–351.

Taboada, M., & Wiesemann, L. (2010). Subjects and topics in conversation. *Journal of Pragmatics*, *42*(7), 1816–1828. https://doi.org/10.1016/j.pragma.2009.04.009

Tarone, E. (1988). *Variation in Interlanguage*. London: Edward Arnold.

Thomas, M. (1994). Assessment of L2 proficiency in Second Language Acquisition research. *Language Learning*, *44*(2), 307–336. https://doi.org/10.1111/j.1467-1770.1994.tb01104.x

Tomlin, R. S. (1987). *Coherence and grounding in discourse*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Tono, Y. (2003). Learner corpora: design, development and applications. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the 2003 Corpus Linguistics Conference* (pp. 800–809). Lancaster: Lancaster University.

Tono, Y. (2004). Multiple comparisons of IL, L1 and TL corpora: the case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and language learners* (pp. 45–66). Amsterdam: John Benjamins.

Toole, J. (1996). The effect of genre on referential choice. In T. Fretheim & J. K. Gundel (Eds.), *Reference and referent accessibility* (pp. 263–290). Amsterdam/Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/pbns.38.16too

Toribio, A. J. (2000). Setting parametric limits on dialectal variation in Spanish. *Lingua*, *10*, 315–341.

Torregrossa, J., & Bongartz, C. (forthcoming). *Activation of referents in the bilingual mind*.

Torregrossa, J., Bongartz, C., & Tsimpli, I. (2015). Testing accessibility: A cross-linguistic comparison of the syntax of referring expressions. *LSA Annual Meeting Extended Abstracts*, *6*, 3–6. https://doi.org/10.3765/exabs.v0i0.3046

Torres Cacoullos, R., & Travis, C. E. (2014). Prosody, priming and particular constructions: the patterning of English first-person singular subject expression in conversation. *Journal of Pragmatics*, *63*, 19–34. https://doi.org/10.1016/j.pragma.2013.08.003

Travis, C. E. (2005). The yo-yo effect: priming in subject expression in Colombian

Spanish. In R. Gess & E. J. Rubin (Eds.), *Theoretical and experimental approaches to Romance Linguistics: selected papers from the 34th Linguistic Symposium on Romance Languages (LSRL)* (pp. 329–349). Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.272.20tra

Travis, C. E. (2007). Genre effects on subject expression in Spanish: priming in narrative and conversation. *Language Variation and Change*, *19*, 101–135. https://doi.org/10.1017/S0954394507070081

Travis, C. E., & Cacoullos, R. T. (2012). What do subject pronouns do in discourse? Cognitive, mechanical and constructional factors in variation. *Cognitive Linguistics*, *23*(4), 711–748. https://doi.org/10.1515/cog-2012-0022

Tsimpli, I., & Dimitrakopoulou, M. (2007). The Interpretability Hypothesis: evidence from wh-interrogatives in second language acquisition. *Second Language Research*, *23*(2), 215–242. https://doi.org/10.1177/0267658307076546

Tsimpli, I., & Roussou, A. (1991). Parameter resetting in L2? *University of Cambridge Working Papers in Linguistics*, *3*, 149–169.

Tsimpli, I., & Sorace, A. (2006). Differentiating interfaces: L2 performance in syntax-semantics and syntax-discourse phenomena. In D. Bamman, T. Magnitskaia, & C. Zaller (Eds.), *Proceedings of the Annual Boston University Conference on Language Development* (pp. 653–664). Somerville MA: Cascadilla Press.

Tsimpli, I., Sorace, A., Heycock, C., & Filiaci, F. (2004). First language attrition and syntactic subjects: a study of Greek and Italian near-native speakers of English. *International Journal of Bilingualism*, *8*(3), 257–277. https://doi.org/10.1177/13670069040080030601

University of Wisconsin. (1998). *The University of Wisconsin College-Level Placement Test: Spanish (Grammar) Form 96M*. Madison, WI: University of Wisconsin Press.

Valenzuela, E. (2006). L2 end state grammars and incomplete acquisition of Spanish CLLD constructions. In R. Slabakova, S. A. Montrul, & P. Prévost (Eds.), *Inquiries in linguistic development* (pp. 283–304). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/z.133.16val

Vallduvi, E. (1992). *The informational component*. New York: Garland.

Van Dijk, T. A. (1977). Sentence topic and discourse topic. *Papers in Slavic Philology*, *1*, 49–61.

Véronique, D., Carroll, M., & von Stutterheim, C. (2000). Anaphoric linkage. Paper presented at the *Euroconference: Information structure, linguistic structure and the dynamics of acquisition.* San Féliu de Guixols.

Vogels, J., Krahmer, E., & Maes, A. (2015). How cognitive load influences speakers' choice of referring expressions. *Cognitive Science*, *39*(6), 1396–1418. https://doi.org/10.1111/cogs.12205

Werth, P. (1984). *Focus, coherence and emphasis*. London: Croom Helm.

Wettstein, H. K. (1984). How to bridge the gap between meaning and reference. *Synthese*, *58*(1), 63–84. https://doi.org/10.1007/BF00485362

White, L. (1985). The pro-drop parameter in adult second language acquisition. *Language Learning*, *35*(1), 47–61. https://doi.org/10.1111/j.1467-1770.1985.tb01014.x

White, L. (1986). Implications of parametric variation for adult second language acquisition: An investigation of the "pro-drop" parameter. In V. Cook (Ed.), *Experimental approaches to second language acquisition* (pp. 55–72). Oxford: Pergamon.

White, L. (1989). *Universal grammar and second language acquisition*. Amsterdam: John Benjamins.

White, L. (2009). Grammatical theory, interfaces and L2 knowledge. In W. Ritchie & T. K. Bhatia (Eds.), *The new handbook of Second Language Acquisition* (pp. 49–70). Bingley, UK: Emeral Press.

White, L. (2011). Second language acquisition at the interfaces. *Lingua*, *121*(4), 577–590. https://doi.org/10.1016/j.lingua.2010.05.005

White, L. (2016). Pro-drop then and now: changing perspectives on null subjects in second language acquisition. In A. A. de la Fuente, E. Valenzuela, & C. M. Sanz (Eds.), *Language acquisition beyond parameters. Studies in honour of Juana M. Liceras* (pp. 17–35). Amsterdam: John Benjamins. https://doi.org/10.1075/sibil.51.02whi

Williams, J. (1988). Zero anaphora in Second Language Acquisition: a comparison among three varieties of English. *Studies in Second Language Acquisition*, *10*(3), 339–370. https://doi.org/10.1017/S0272263100007488

Willis, J. (1998). Concordances in the classroom without a computer: assembling and

exploiting concordances of common words. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp. 44–66). Cambridge: Cambridge University Press.

Zhao, L. X. (2014). Ultimate attainment of anaphora resolution in L2 Chinese. *Second Language Research*, *30*(3), 381–407. https://doi.org/10.1177/0267658314521107

Zulaica-Hernández, I. (2016). Topic-continuity and topic-shift effects in Spanish discourse. *International Review of Pragmatics*, *8*(1), 1–35. https://doi.org/10.1163/18773109-00801001

APPENDICES

A. Participants: basic biodata and learning background forms

Figure 48. Basic biodata and learning background form (Spanish L1)

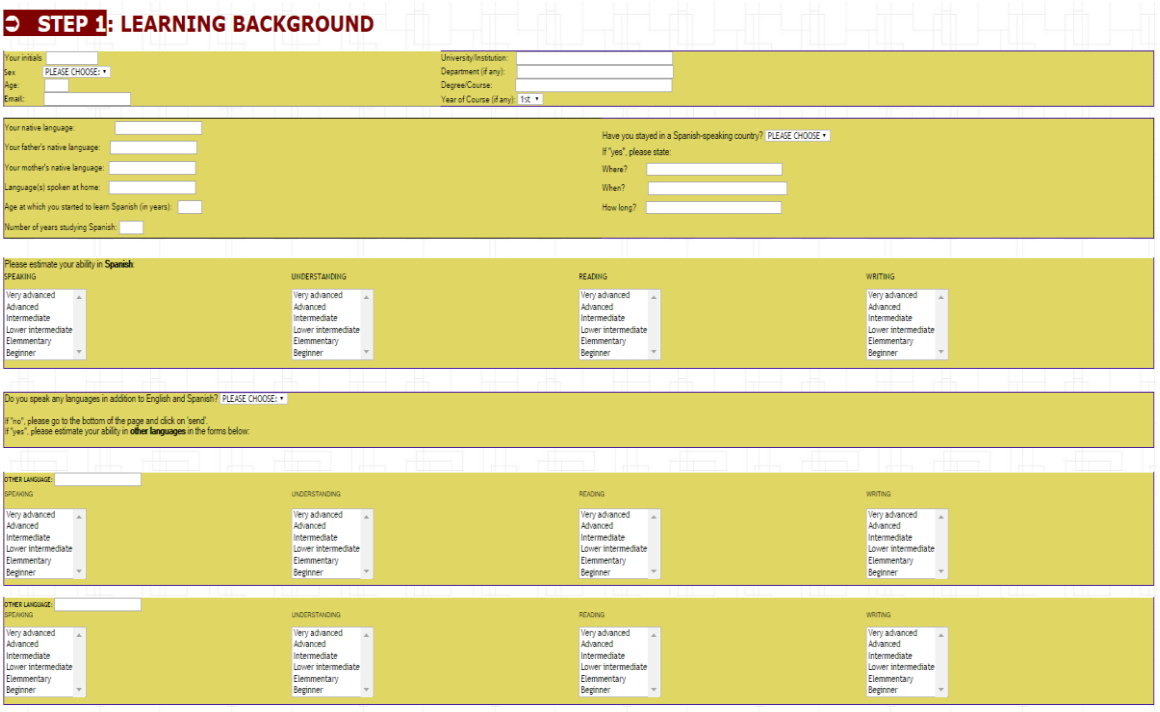


Figure 49. Basic biodata and learning background form (English L1)

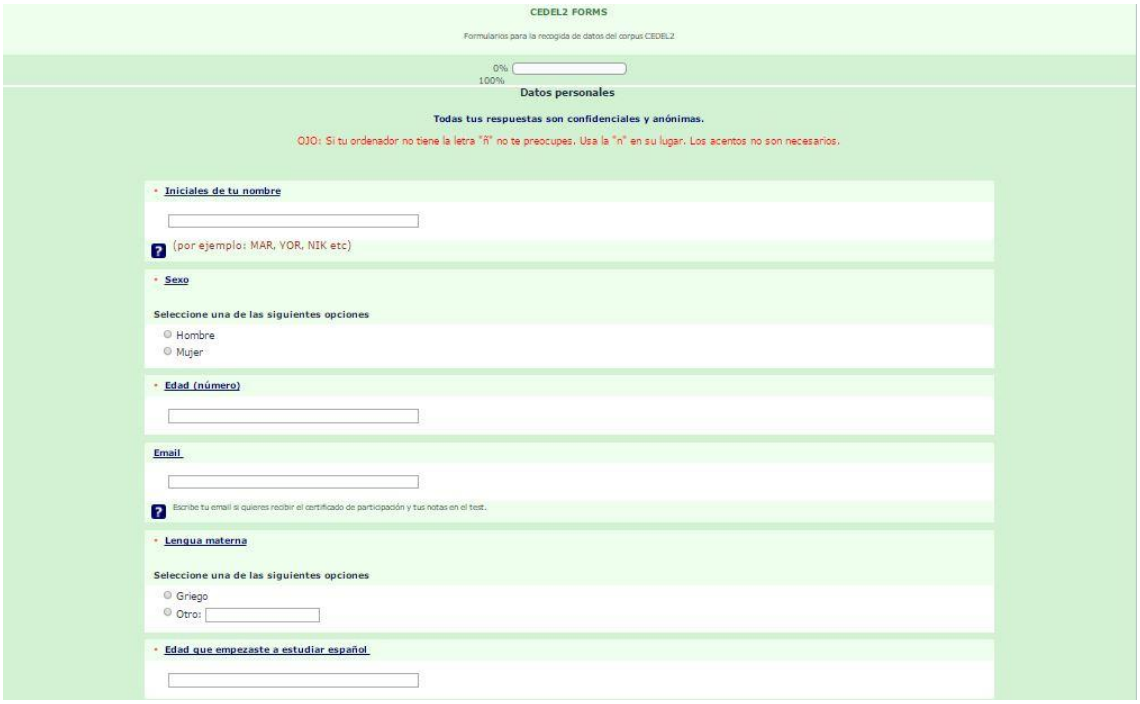
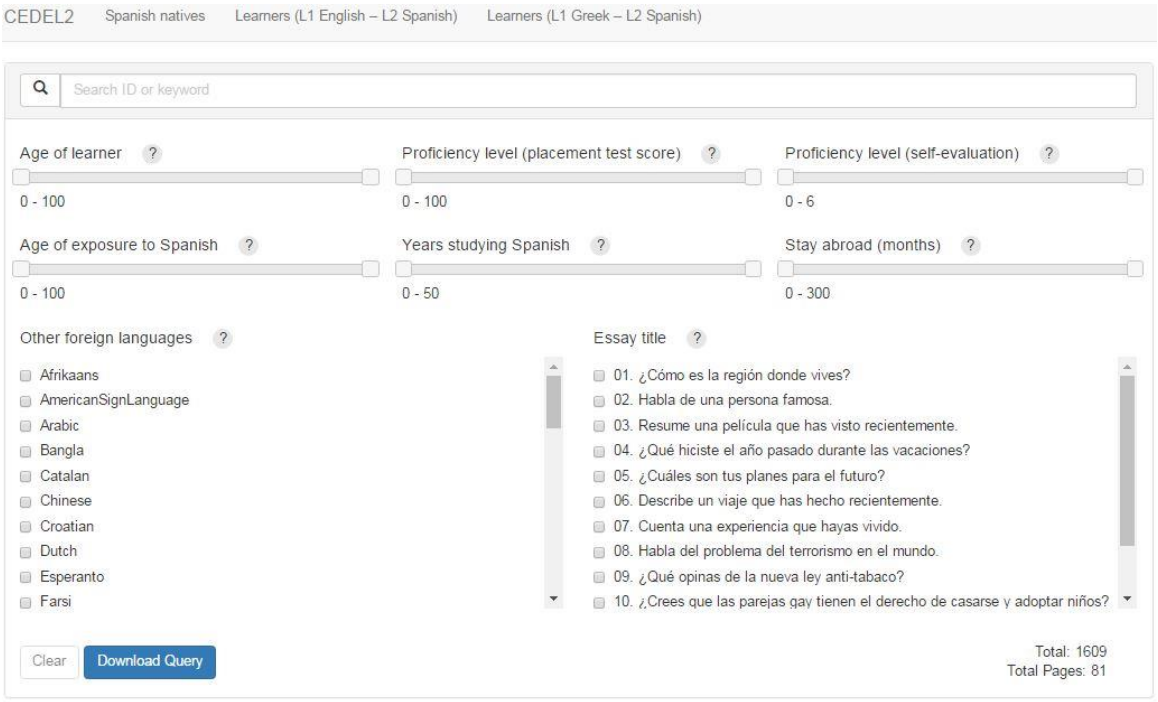


Figure 50. Basic biodata and learning background form (Greek L1)

B. CEDEL2 interface and UAM corpus tool annotation schemes



ID	Age of learner	Proficiency level (placement test score)	Proficiency level (self-evaluation)	Age of exposure to Spanish	Years studying Spanish	Stay abroad (months)	Other foreign languages	Essay title
9_26_0_2_nlp	26	21	1.75	26	0			02. Habla de una persona famosa.
9_25_0_5_jw	25	21	1.5	25	0			05. ¿Cuáles son tus planes para el futuro?
9_23_4_5_jar	23	21	1	16	4			05. ¿Cuáles son tus planes para el futuro?
9_20_4_5_mls	20	21	2.25	16	4			05. ¿Cuáles son tus planes para el futuro?

Figure 51. CEDEL2 corpus online query interface

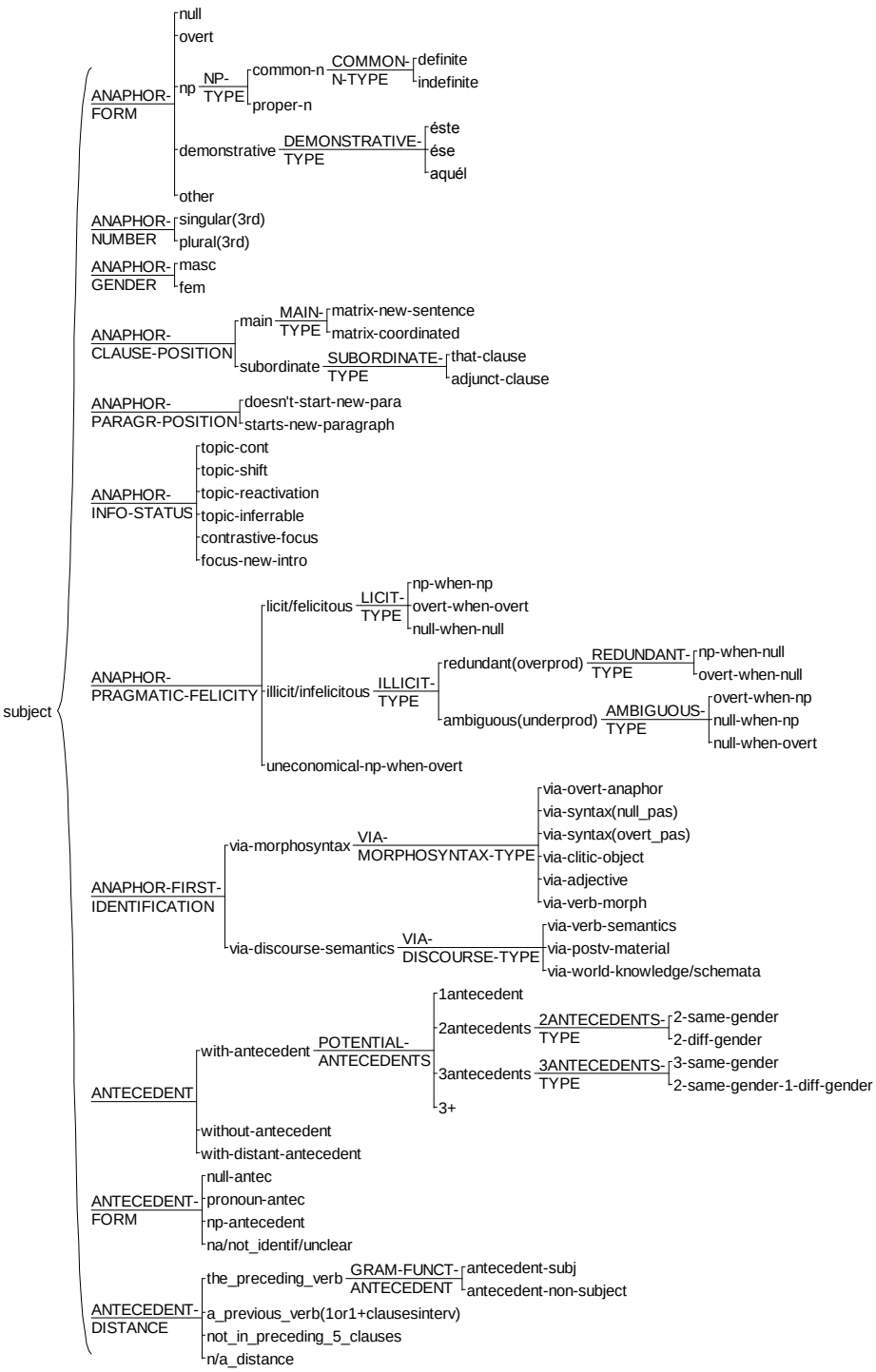


Figure 52. Lozano's annotation scheme (Lozano, 2016:251)

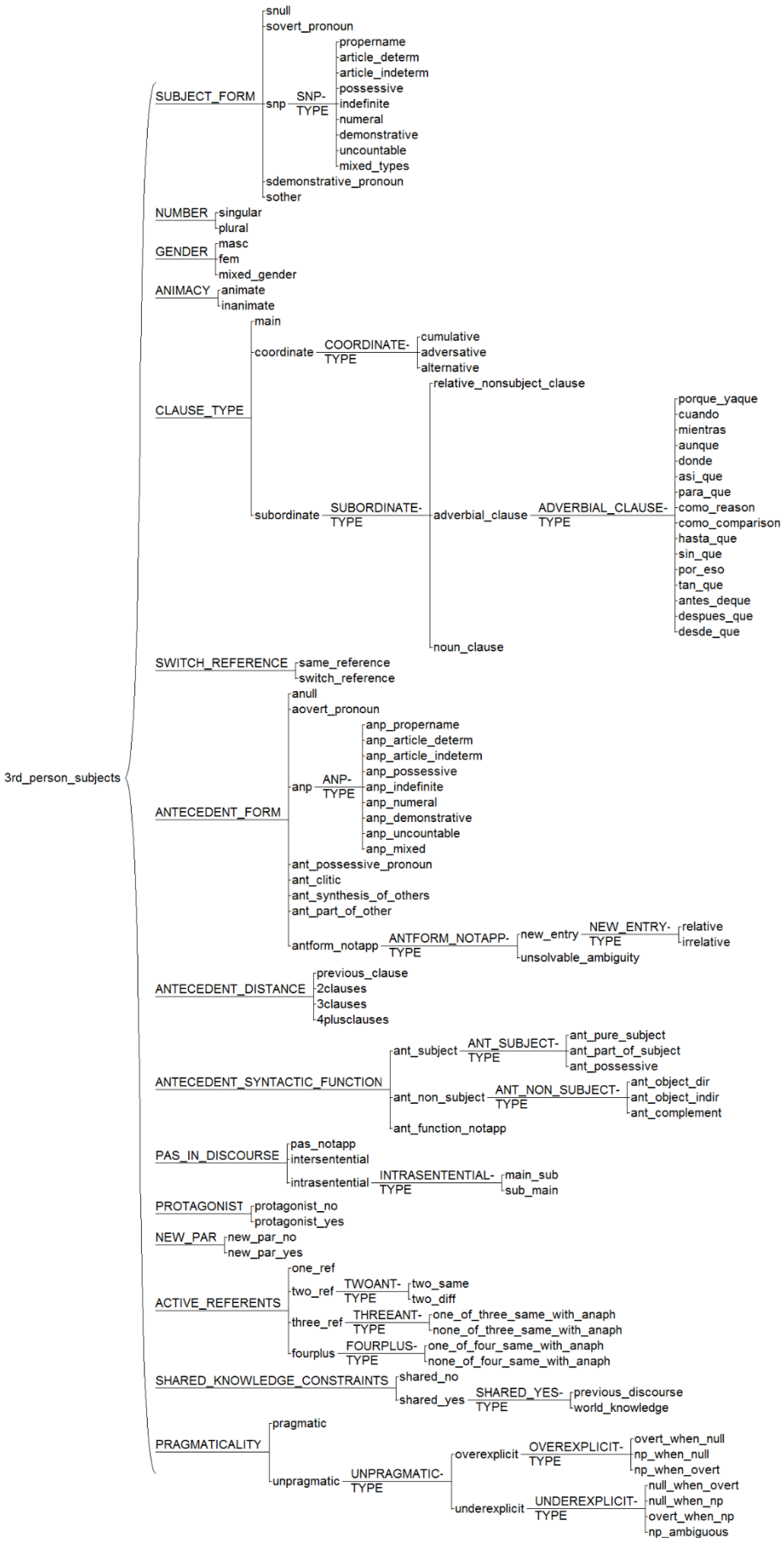


Figure 53. The annotation scheme of the present study

C. Results: original search queries and raw frequencies

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

Unit: 3rd_person_subjects + | Show

Set 1: natives + | Set 2: english1 + | Set 3: english2 + | Set 4: english3 + | Set 5: greek1 + | Set 6: greek2 + | Set 7: greek3 + | +

	natives		english1		english2		english3		greek1		greek2		greek3	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Total Units	368		323		281		425		202		248		213	
SUBJECT_FORM	N=368		N=323		N=281		N=425		N=202		N=248		N=213	
- snull	219	59.51%	82	25.39%	117	41.64%	209	49.18%	141	69.80%	172	69.35%	133	62.44%
- sovert_pronoun	26	7.07%	105	32.51%	55	19.57%	65	15.29%	16	7.92%	11	4.44%	9	4.23%
- snp	121	32.88%	134	41.49%	107	38.08%	149	35.06%	45	22.28%	63	25.40%	71	33.33%
- sdemonstrative_pron	2	0.54%	0	0.00%	1	0.36%	1	0.24%	0	0.00%	0	0.00%	0	0.00%
- sother	0	0.00%	2	0.62%	1	0.36%	1	0.24%	0	0.00%	2	0.81%	0	0.00%

Figure 54. Overall distribution of forms

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

Unit: main + | Show

Set 1: natives + | +

Feature	N	Percent
Total Units	176	
SUBJECT_FORM	N=176	
- snull	72	40.91%
- sovert	104	59.09%

Figure 55. Main clauses (native speakers)

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

Unit: coordinate + | Show

Set 1: natives + | +

Feature	N	Percent
Total Units	78	
SUBJECT_FORM	N=78	
- snull	73	93.59%
- sovert	5	6.41%

Figure 56. Coordinate clauses (native speakers)

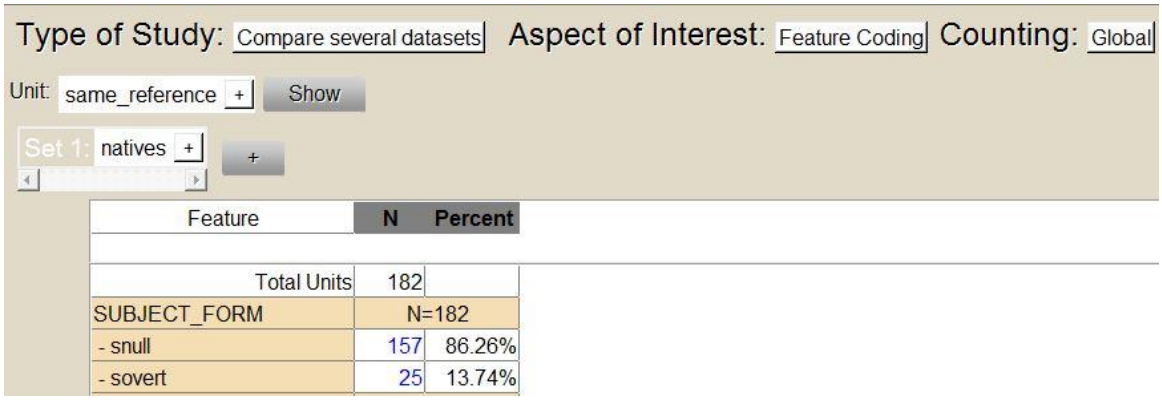
Type of Study: Compare several datasets | Aspect of Interest: Feature Coding | Counting: Global

Unit: subordinate + | Show

Set 1: natives + | +

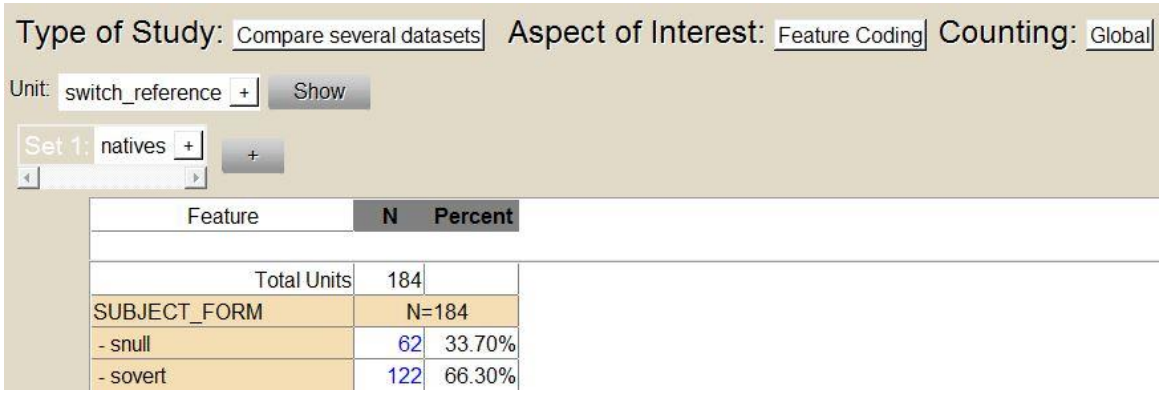
Feature	N	Percent
Total Units	112	
SUBJECT_FORM	N=112	
- snull	74	66.07%
- sovert	38	33.93%

Figure 57. Subordinate clauses (native speakers)



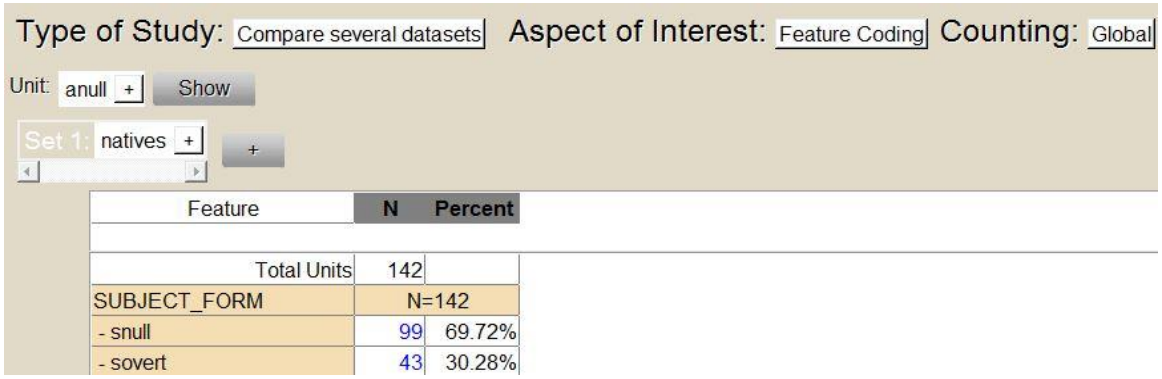
Feature	N	Percent
Total Units	182	
SUBJECT_FORM	N=182	
- snull	157	86.26%
- sovert	25	13.74%

Figure 58. Same-reference (native speakers)



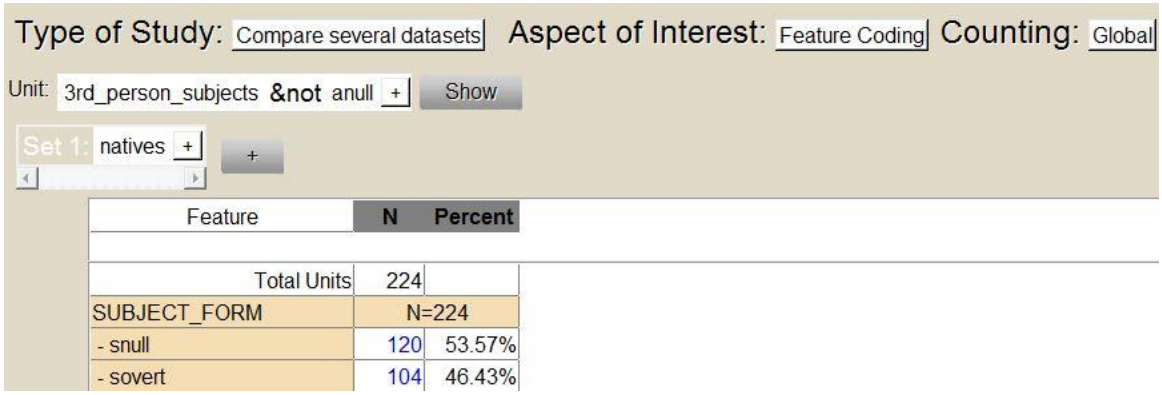
Feature	N	Percent
Total Units	184	
SUBJECT_FORM	N=184	
- snull	62	33.70%
- sovert	122	66.30%

Figure 59. Switch-reference (native speakers)



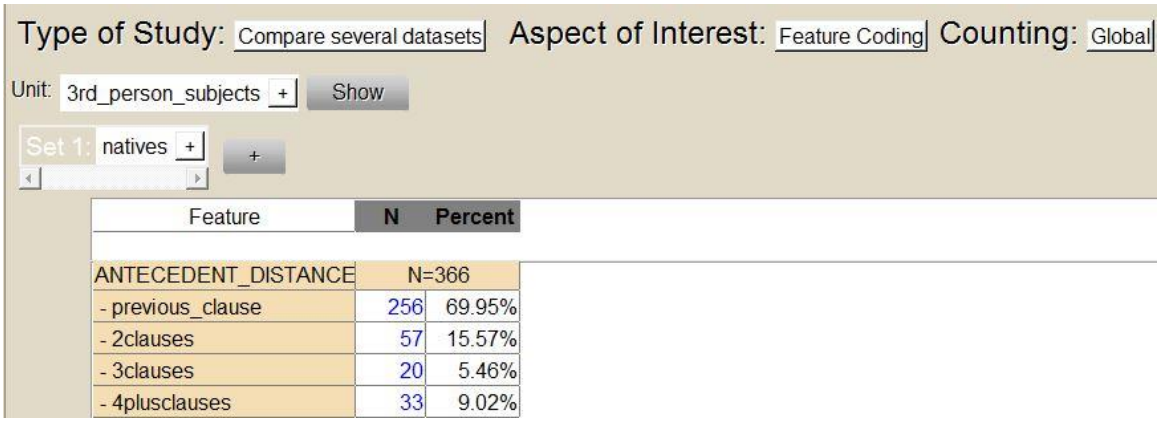
Feature	N	Percent
Total Units	142	
SUBJECT_FORM	N=142	
- snull	99	69.72%
- sovert	43	30.28%

Figure 60. Null antecedent (native speakers)



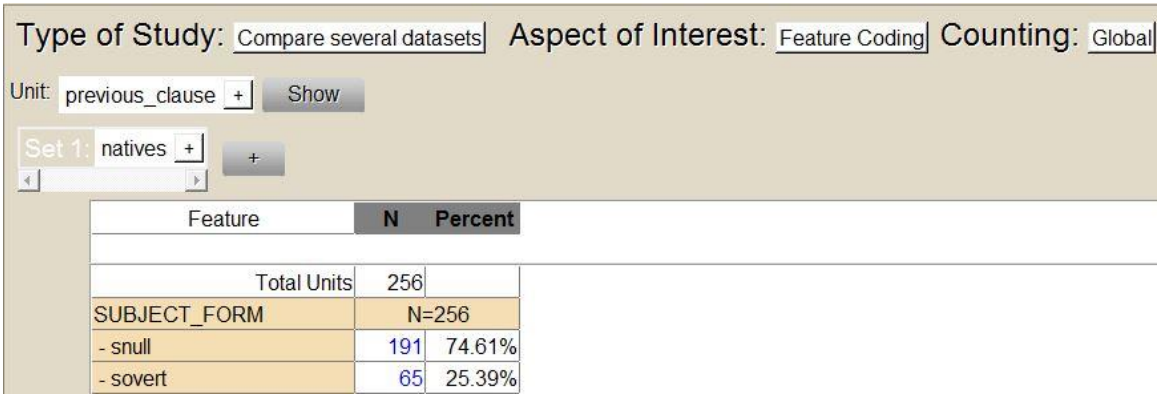
Feature	N	Percent
Total Units	224	
SUBJECT_FORM	N=224	
- snull	120	53.57%
- sovert	104	46.43%

Figure 61. Overt antecedent (native speakers)



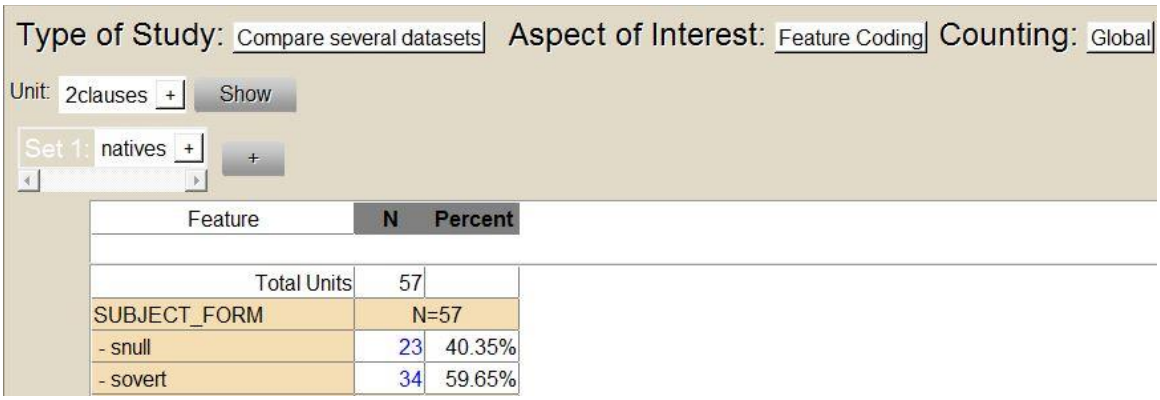
Feature	N	Percent
ANTECEDENT_DISTANCE	N=366	
- previous_clause	256	69.95%
- 2clauses	57	15.57%
- 3clauses	20	5.46%
- 4plusclauses	33	9.02%

Figure 62. Anaphors per antecedent distance (natives)



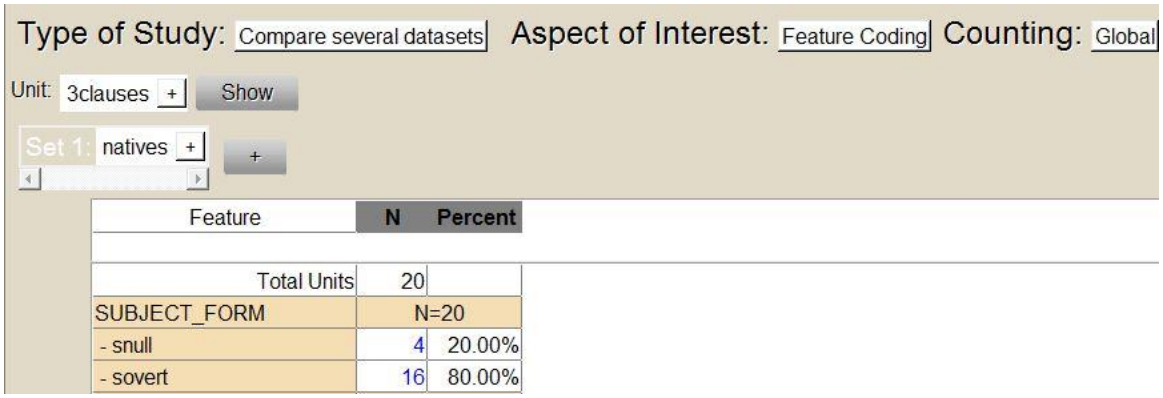
Feature	N	Percent
Total Units	256	
SUBJECT_FORM	N=256	
- snull	191	74.61%
- sovert	65	25.39%

Figure 63. One clause distance (natives)



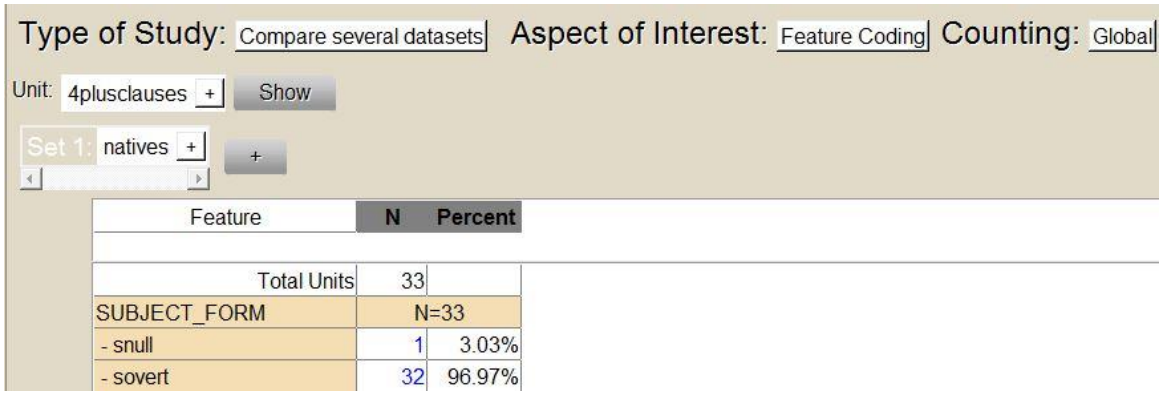
Feature	N	Percent
Total Units	57	
SUBJECT_FORM	N=57	
- snull	23	40.35%
- sovert	34	59.65%

Figure 64. Two clauses distance (natives)



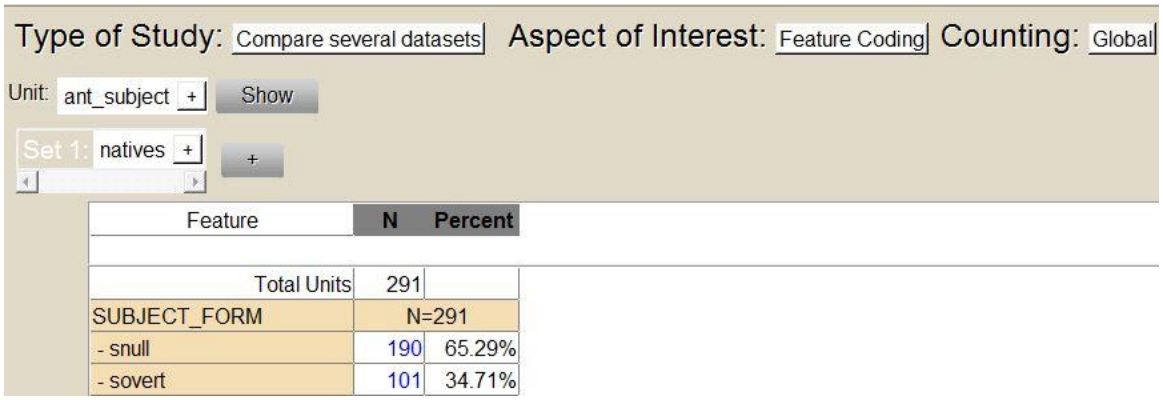
Feature	N	Percent
Total Units	20	
SUBJECT_FORM	N=20	
- snull	4	20.00%
- sovert	16	80.00%

Figure 65. Three clauses distance (natives)



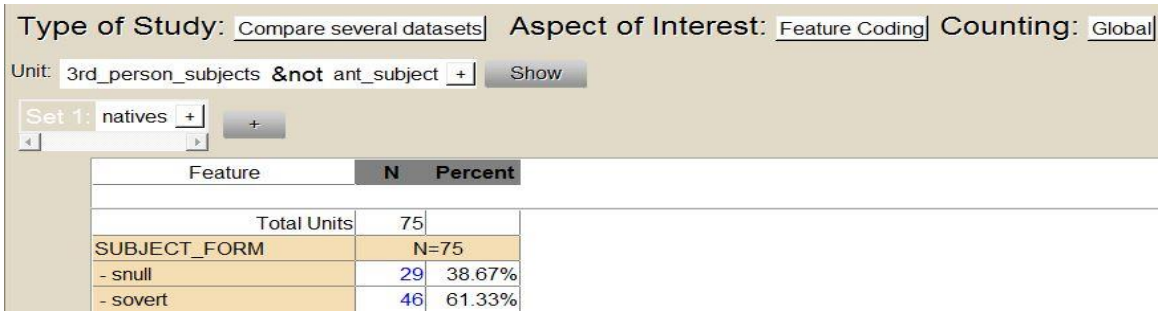
Feature	N	Percent
Total Units	33	
SUBJECT_FORM	N=33	
- snull	1	3.03%
- sovert	32	96.97%

Figure 66. Four (+) clauses distance (natives)



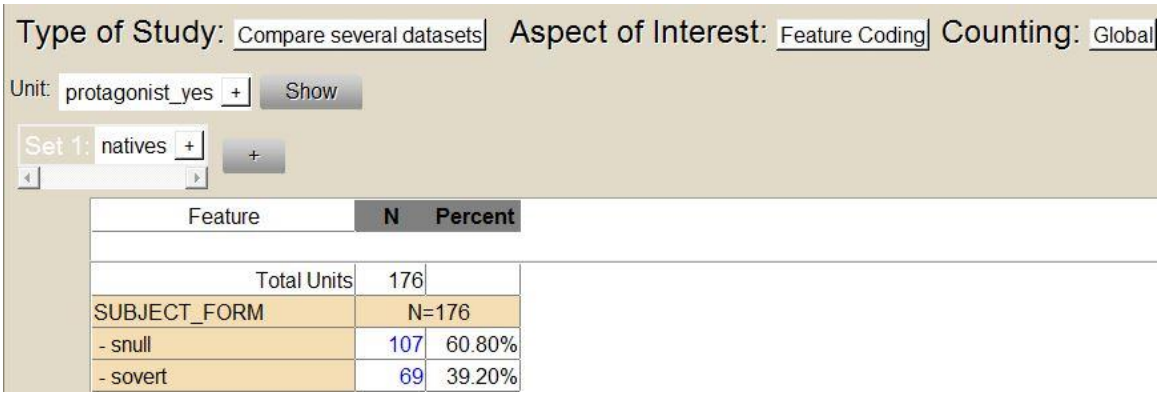
Feature	N	Percent
Total Units	291	
SUBJECT_FORM	N=291	
- snull	190	65.29%
- sovert	101	34.71%

Figure 67. Subject antecedent (natives)



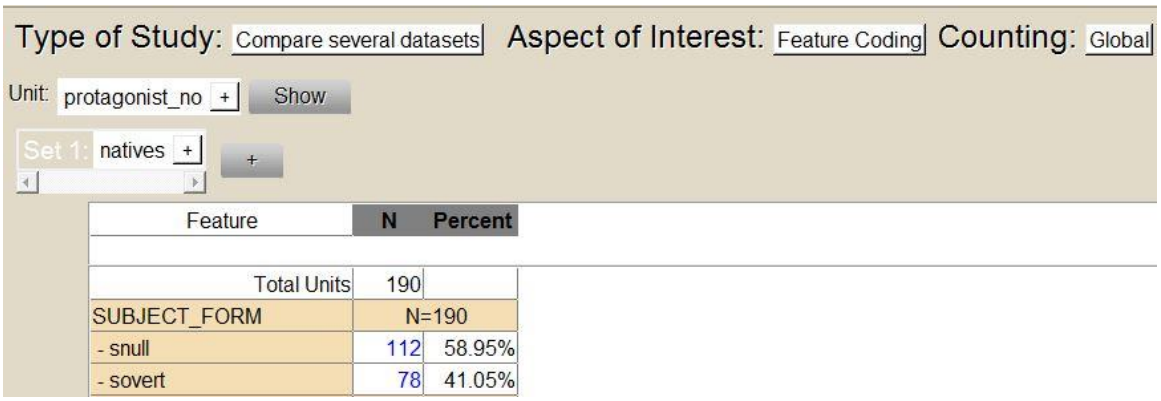
Feature	N	Percent
Total Units	75	
SUBJECT_FORM	N=75	
- snull	29	38.67%
- sovert	46	61.33%

Figure 68. Non-subject antecedent (natives)



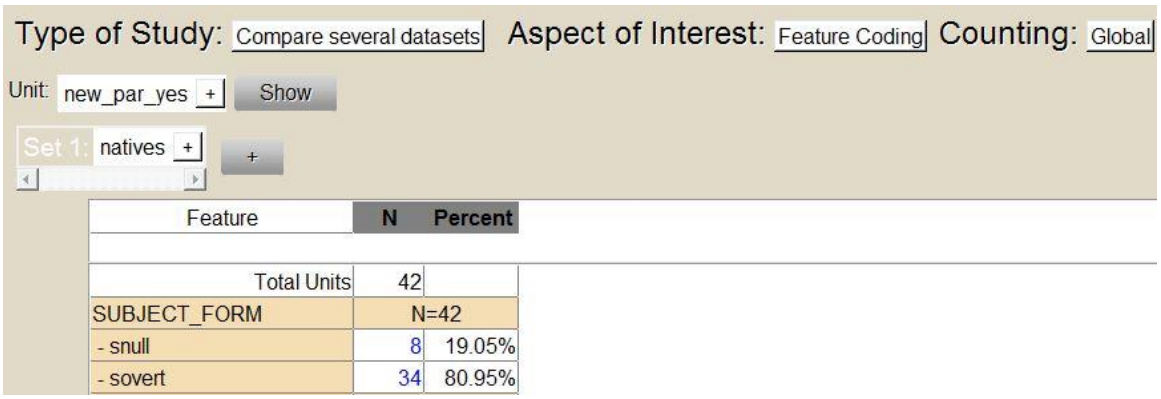
Feature	N	Percent
Total Units	176	
SUBJECT_FORM	N=176	
- snull	107	60.80%
- sovert	69	39.20%

Figure 69. Protagonist antecedent (natives)



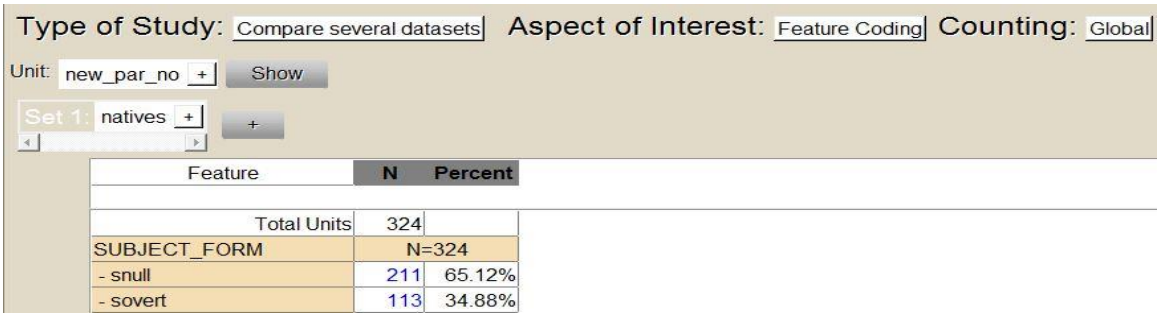
Feature	N	Percent
Total Units	190	
SUBJECT_FORM	N=190	
- snull	112	58.95%
- sovert	78	41.05%

Figure 70. Non-protagonist antecedent (natives)



Feature	N	Percent
Total Units	42	
SUBJECT_FORM	N=42	
- snull	8	19.05%
- sovert	34	80.95%

Figure 71. New paragraph (natives)



Feature	N	Percent
Total Units	324	
SUBJECT_FORM	N=324	
- snull	211	65.12%
- sovert	113	34.88%

Figure 72. Same paragraph (natives)

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

Unit: one_ref + | Show

Set 1: natives + | +

Feature	N	Percent
Total Units	206	
SUBJECT_FORM	N=206	
- snull	151	73.30%
- sovert	55	26.70%

Figure 73. One active referent (natives)

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

Unit: two_ref + | Show

Set 1: natives + | +

Feature	N	Percent
Total Units	102	
SUBJECT_FORM	N=102	
- snull	53	51.96%
- sovert	49	48.04%

Figure 74. Two active referents (natives)

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

Unit: three_ref + | Show

Set 1: natives + | +

Feature	N	Percent
Total Units	40	
SUBJECT_FORM	N=40	
- snull	15	37.50%
- sovert	25	62.50%

Figure 75. Three active referents (natives)

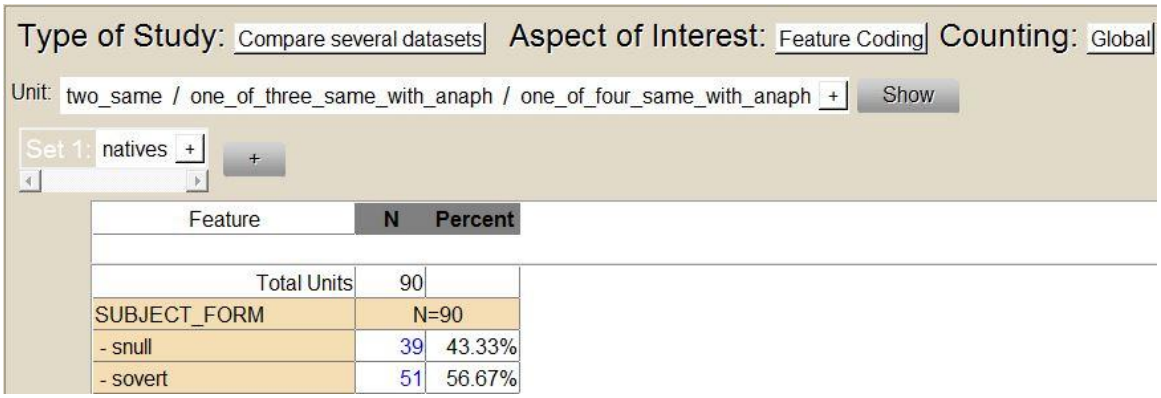
Type of Study: Compare several datasets | Aspect of Interest: Feature Coding | Counting: Global

Unit: fourplus + | Show

Set 1: natives + | +

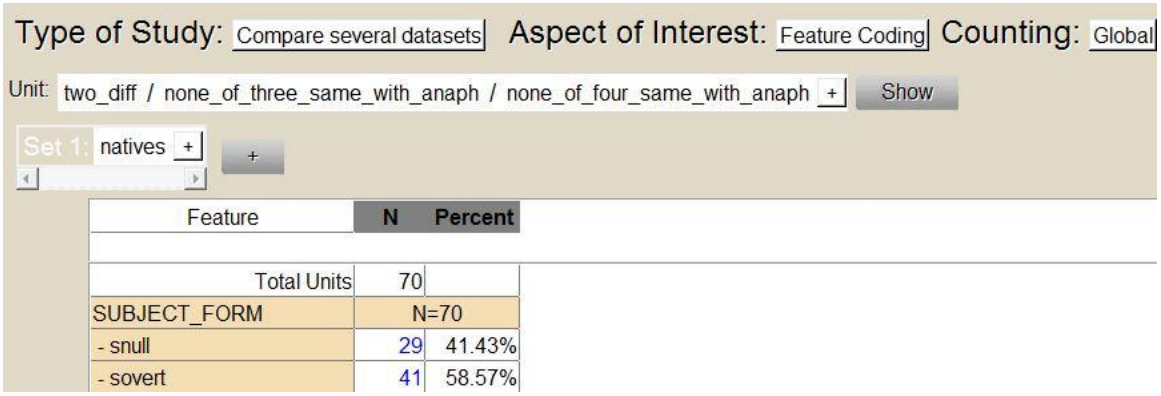
Feature	N	Percent
Total Units	18	
SUBJECT_FORM	N=18	
- snull	0	0.00%
- sovert	18	100.00%

Figure 76. Four (+) active referents (natives)



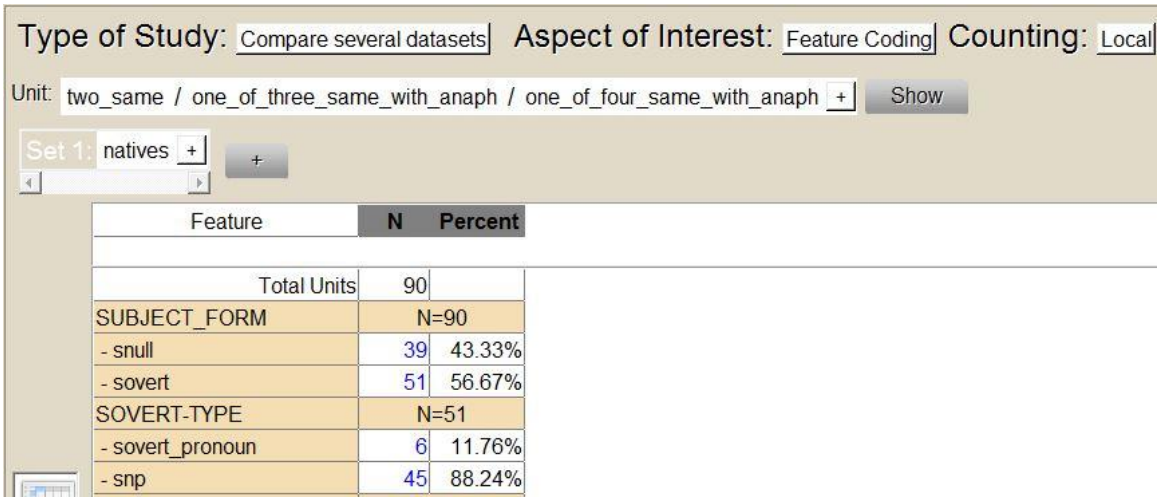
Feature	N	Percent
Total Units	90	
SUBJECT_FORM	N=90	
- snull	39	43.33%
- sovert	51	56.67%

Figure 77. Same gender referents (natives)



Feature	N	Percent
Total Units	70	
SUBJECT_FORM	N=70	
- snull	29	41.43%
- sovert	41	58.57%

Figure 78. Different gender referents (natives)



Feature	N	Percent
Total Units	90	
SUBJECT_FORM	N=90	
- snull	39	43.33%
- sovert	51	56.67%
SOVERT-TYPE	N=51	
- sovert_pronoun	6	11.76%
- snp	45	88.24%

Figure 79. Same gender referents by overt type (natives)

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

Unit: two_diff / none_of_three_same_with_anaph / none_of_four_same_with_anaph + | Show

Set 1: natives + | +

Feature	N	Percent
Total Units	70	
SUBJECT_FORM	N=70	
- snull	29	41.43%
- sovert	41	58.57%
SOVERT-TYPE	N=41	
- sovert_pronoun	14	34.15%
- snp	27	65.85%

Figure 80. Different gender referents by overt type (natives)

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

Unit: shared_no + | Show

Set 1: natives + | +

Feature	N	Percent
Total Units	298	
SUBJECT_FORM	N=298	
- snull	175	58.72%
- sovert	123	41.28%

Figure 81. No shared knowledge (natives)

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

Unit: shared_yes + | Show

Set 1: natives + | +

Feature	N	Percent
Total Units	68	
SUBJECT_FORM	N=68	
- snull	44	64.71%
- sovert	24	35.29%

Figure 82. Shared knowledge (natives)

Type of Study: Describe a dataset | Aspect of Interest: Feature Coding | Counting: Local

Unit: 3rd_person_subjects + | Show

Feature	N	Percent
PRAGMATICALITY	N=2050	
- pragmatic	1572	76.68%
- unpragmatic	478	23.32%
UNPRAGMATIC-TYPE	N=478	
- overexplicit	412	86.19%
- underexplicit	66	13.81%
OVEREXPLICIT-TYPE	N=412	
- overt_when_null	210	50.97%
- np_when_null	139	33.74%
- np_when_overt	63	15.29%
UNDEREXPLICIT-TYPE	N=66	
- null_when_overt	20	30.30%
- null_when_np	31	46.97%
- overt_when_np	10	15.15%
- np_ambiguous	5	7.58%

Figure 83. Pragmaticality (all groups together)

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

Unit: 3rd_person_subjects + | Show

Set 1: english1 + | Set 2: greek1 + | Set 3: english2 + | Set 4: greek2 + | Set 5: english3 + | Set 6: greek3 + | Set 7: natives + | +

	english1		greek1		english2		greek2		english3		greek3		natives	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=321		N=202		N=279		N=246		N=423		N=213		N=366	
- pragmatic	169	52.65%	168	83.17%	195	69.89%	212	86.18%	323	76.36%	190	89.20%	313	85.52%
- unpragmatic	152	47.35%	34	16.83%	84	30.11%	34	13.82%	100	23.64%	23	10.80%	53	14.48%
UNPRAGMATIC-TYPE	N=152		N=34		N=84		N=34		N=100		N=23		N=53	
- overexplicit	147	96.71%	32	94.12%	79	94.05%	30	88.24%	85	85.00%	15	65.22%	24	45.28%
- underexplicit	5	3.29%	2	5.88%	5	5.95%	4	11.76%	15	15.00%	8	34.78%	29	54.72%
OVEREXPLICIT-TYPE	N=147		N=32		N=79		N=30		N=85		N=15		N=24	
- overt_when_null	91	61.90%	17	53.12%	49	62.03%	8	26.67%	39	45.88%	2	13.33%	4	16.67%
- np_when_null	40	27.21%	14	43.75%	27	34.18%	16	53.33%	25	29.41%	7	46.67%	9	37.50%
- np_when_overt	16	10.88%	1	3.12%	3	3.80%	6	20.00%	21	24.71%	6	40.00%	11	45.83%
UNDEREXPLICIT-TYPE	N=5		N=2		N=5		N=4		N=15		N=8		N=29	
- null_when_overt	1	20.00%	0	0.00%	1	20.00%	2	50.00%	5	33.33%	1	12.50%	11	37.93%
- null_when_np	1	20.00%	2	100.00%	2	40.00%	1	25.00%	7	46.67%	4	50.00%	15	51.72%
- overt_when_np	3	60.00%	0	0.00%	1	20.00%	0	0.00%	2	13.33%	2	25.00%	2	6.90%
- np_ambiguous	0	0.00%	0	0.00%	1	20.00%	1	25.00%	1	6.67%	1	12.50%	1	3.45%

Figure 84. Pragmaticality per group

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

Unit: 3rd_person_subjects + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 + Set 7: natives +

	english1		greek1		english2		greek2		english3		greek3		natives	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=316		N=200		N=274		N=242		N=408		N=205		N=337	
- pragmatic	169	53.48%	168	84.00%	195	71.17%	212	87.60%	323	79.17%	190	92.68%	313	92.88%
- unpragmatic	147	46.52%	32	16.00%	79	28.83%	30	12.40%	85	20.83%	15	7.32%	24	7.12%
UNPRAGMATIC-TYPE	N=147		N=32		N=79		N=30		N=85		N=15		N=24	
- overexplicit	147	100.00%	32	100.00%	79	100.00%	30	100.00%	85	100.00%	15	100.00%	24	100.00%
- underexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
OVEREXPLICIT-TYPE	N=147		N=32		N=79		N=30		N=85		N=15		N=24	
- overt_when_null	91	61.90%	17	53.12%	49	62.03%	8	26.67%	39	45.88%	2	13.33%	4	16.67%
- np_when_null	40	27.21%	14	43.75%	27	34.18%	16	53.33%	25	29.41%	7	46.67%	9	37.50%
- np_when_overt	16	10.88%	1	3.12%	3	3.80%	6	20.00%	21	24.71%	6	40.00%	11	45.83%
UNDEREXPLICIT-TYPE	N=0		N=0		N=0		N=0		N=0		N=0		N=0	
- null_when_overt	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
- null_when_np	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
- overt_when_np	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
- np_ambiguous	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 85. Overexplicitness per group

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

Unit: sing + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 +

	english1		greek1		english2		greek2		english3		greek3	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=274		N=195		N=248		N=218		N=364		N=167	
- pragmatic	143	52.19%	163	83.59%	175	70.56%	190	87.16%	282	77.47%	154	92.22%
- unpragmatic	131	47.81%	32	16.41%	73	29.44%	28	12.84%	82	22.53%	13	7.78%
UNPRAGMATIC-TYPE	N=131		N=32		N=73		N=28		N=82		N=13	
- overexplicit	131	100.00%	32	100.00%	73	100.00%	28	100.00%	82	100.00%	13	100.00%
- underexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 86. Overexplicitness in singular number

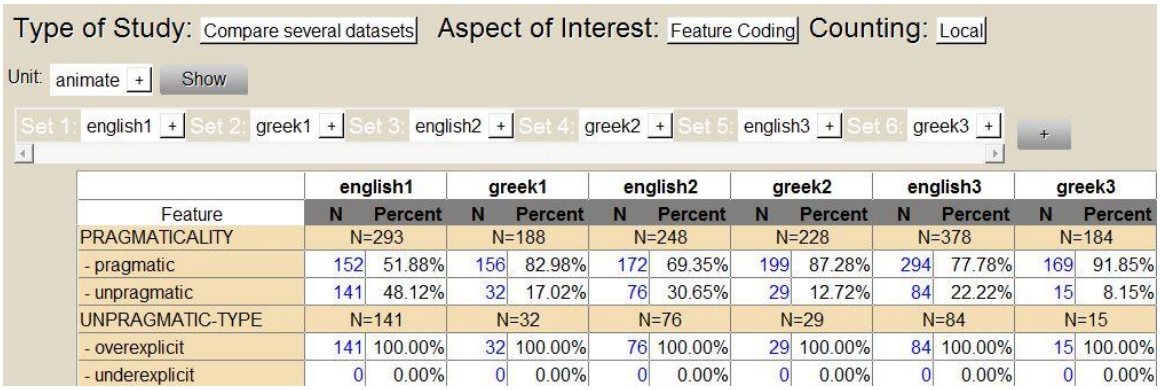
Type of Study: Compare several datasets Aspect of Interest: Feature Coding Counting: Local

Unit: pl + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 +

	english1		greek1		english2		greek2		english3		greek3	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=42		N=5		N=26		N=24		N=44		N=38	
- pragmatic	26	61.90%	5	100.00%	20	76.92%	22	91.67%	41	93.18%	36	94.74%
- unpragmatic	16	38.10%	0	0.00%	6	23.08%	2	8.33%	3	6.82%	2	5.26%
UNPRAGMATIC-TYPE	N=16		N=0		N=6		N=2		N=3		N=2	
- overexplicit	16	100.00%	0	0.00%	6	100.00%	2	100.00%	3	100.00%	2	100.00%
- underexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 87. Overexplicitness in plural number

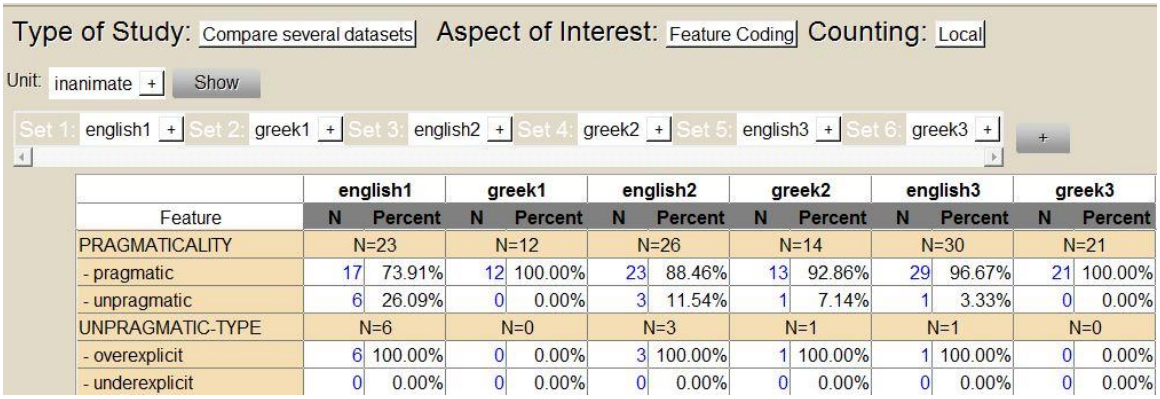

Type of Study: Compare several datasets Aspect of Interest: Feature Coding Counting: Local

Unit: animate + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 +

	english1		greek1		english2		greek2		english3		greek3	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=293		N=188		N=248		N=228		N=378		N=184	
- pragmatic	152	51.88%	156	82.98%	172	69.35%	199	87.28%	294	77.78%	169	91.85%
- unpragmatic	141	48.12%	32	17.02%	76	30.65%	29	12.72%	84	22.22%	15	8.15%
UNPRAGMATIC-TYPE	N=141		N=32		N=76		N=29		N=84		N=15	
- overexplicit	141	100.00%	32	100.00%	76	100.00%	29	100.00%	84	100.00%	15	100.00%
- underexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 88. Overexplicitness with animate subjects

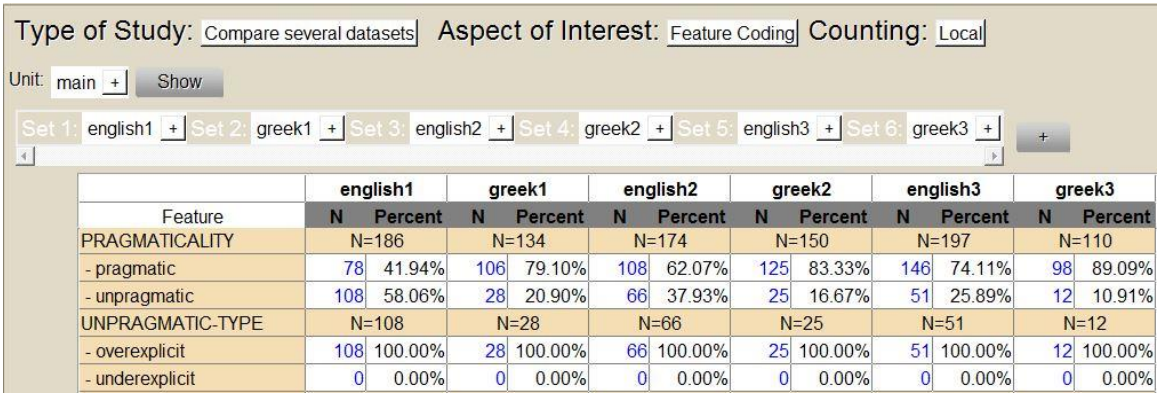

Type of Study: Compare several datasets Aspect of Interest: Feature Coding Counting: Local

Unit: inanimate + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 +

	english1		greek1		english2		greek2		english3		greek3	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=23		N=12		N=26		N=14		N=30		N=21	
- pragmatic	17	73.91%	12	100.00%	23	88.46%	13	92.86%	29	96.67%	21	100.00%
- unpragmatic	6	26.09%	0	0.00%	3	11.54%	1	7.14%	1	3.33%	0	0.00%
UNPRAGMATIC-TYPE	N=6		N=0		N=3		N=1		N=1		N=0	
- overexplicit	6	100.00%	0	0.00%	3	100.00%	1	100.00%	1	100.00%	0	0.00%
- underexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 89. Overexplicitness with inanimate subjects

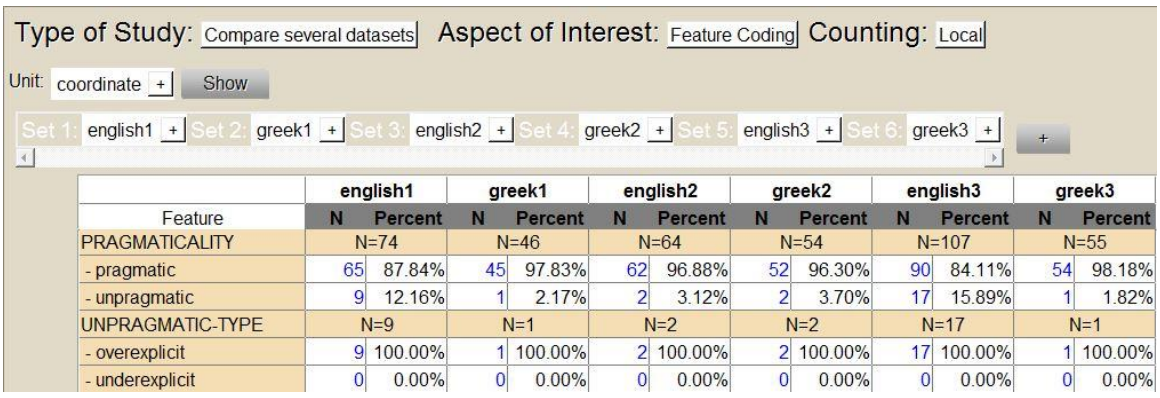

Type of Study: Compare several datasets Aspect of Interest: Feature Coding Counting: Local

Unit: main + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 +

	english1		greek1		english2		greek2		english3		greek3	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=186		N=134		N=174		N=150		N=197		N=110	
- pragmatic	78	41.94%	106	79.10%	108	62.07%	125	83.33%	146	74.11%	98	89.09%
- unpragmatic	108	58.06%	28	20.90%	66	37.93%	25	16.67%	51	25.89%	12	10.91%
UNPRAGMATIC-TYPE	N=108		N=28		N=66		N=25		N=51		N=12	
- overexplicit	108	100.00%	28	100.00%	66	100.00%	25	100.00%	51	100.00%	12	100.00%
- underexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 90. Overexplicitness in main clauses


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

Unit: coordinate + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 +

	english1		greek1		english2		greek2		english3		greek3	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=74		N=46		N=64		N=54		N=107		N=55	
- pragmatic	65	87.84%	45	97.83%	62	96.88%	52	96.30%	90	84.11%	54	98.18%
- unpragmatic	9	12.16%	1	2.17%	2	3.12%	2	3.70%	17	15.89%	1	1.82%
UNPRAGMATIC-TYPE	N=9		N=1		N=2		N=2		N=17		N=1	
- overexplicit	9	100.00%	1	100.00%	2	100.00%	2	100.00%	17	100.00%	1	100.00%
- underexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 91. Overexplicitness in coordinate clauses

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

Unit: subordinate + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 +

	english1		greek1		english2		greek2		english3		greek3	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=56		N=20		N=36		N=38		N=104		N=40	
- pragmatic	26	46.43%	17	85.00%	25	69.44%	35	92.11%	87	83.65%	38	95.00%
- unpragmatic	30	53.57%	3	15.00%	11	30.56%	3	7.89%	17	16.35%	2	5.00%
UNPRAGMATIC-TYPE	N=30		N=3		N=11		N=3		N=17		N=2	
- overexplicit	30	100.00%	3	100.00%	11	100.00%	3	100.00%	17	100.00%	2	100.00%
- underexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 92. Overexplicitness in subordinate clauses

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

Unit: coordinate & same_reference + Show

Set 1: english1 + Set 2: english2 + Set 3: english3 +

	english1		english2		english3	
Feature	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=54		N=51		N=78	
- pragmatic	51	94.44%	49	96.08%	72	92.31%
- unpragmatic	3	5.56%	2	3.92%	6	7.69%

Figure 93. Overexplicitness in same-reference coordinate clauses

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

Unit: one_ref + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 +

	english1		greek1		english2		greek2		english3		greek3	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=194		N=166		N=174		N=177		N=205		N=112	
- pragmatic	82	42.27%	138	83.13%	111	63.79%	157	88.70%	166	80.98%	104	92.86%
- unpragmatic	112	57.73%	28	16.87%	63	36.21%	20	11.30%	39	19.02%	8	7.14%
UNPRAGMATIC-TYPE	N=112		N=28		N=63		N=20		N=39		N=8	
- overexplicit	112	100.00%	28	100.00%	63	100.00%	20	100.00%	39	100.00%	8	100.00%
- underexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 94. Overexplicitness with one active referent

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

Unit: two_ref + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 +

	english1		greek1		english2		greek2		english3		greek3	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=77		N=17		N=66		N=48		N=135		N=47	
- pragmatic	49	63.64%	15	88.24%	55	83.33%	41	85.42%	96	71.11%	42	89.36%
- unpragmatic	28	36.36%	2	11.76%	11	16.67%	7	14.58%	39	28.89%	5	10.64%
UNPRAGMATIC-TYPE	N=28		N=2		N=11		N=7		N=39		N=5	
- overexplicit	28	100.00%	2	100.00%	11	100.00%	7	100.00%	39	100.00%	5	100.00%
- underexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 95. Overexplicitness with two active referents

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

Unit: three_ref + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 +

	english1		greek1		english2		greek2		english3		greek3	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=25		N=13		N=19		N=14		N=41		N=26	
- pragmatic	19	76.00%	11	84.62%	15	78.95%	11	78.57%	34	82.93%	24	92.31%
- unpragmatic	6	24.00%	2	15.38%	4	21.05%	3	21.43%	7	17.07%	2	7.69%
UNPRAGMATIC-TYPE	N=6		N=2		N=4		N=3		N=7		N=2	
- overexplicit	6	100.00%	2	100.00%	4	100.00%	3	100.00%	7	100.00%	2	100.00%
- underexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 96. Overexplicitness with three active referents

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

Unit: fourplus + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 +

	english1		greek1		english2		greek2		english3		greek3	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=20		N=4		N=15		N=3		N=27		N=20	
- pragmatic	19	95.00%	4	100.00%	14	93.33%	3	100.00%	27	100.00%	20	100.00%
- unpragmatic	1	5.00%	0	0.00%	1	6.67%	0	0.00%	0	0.00%	0	0.00%
UNPRAGMATIC-TYPE	N=1		N=0		N=1		N=0		N=0		N=0	
- overexplicit	1	100.00%	0	0.00%	1	100.00%	0	0.00%	0	0.00%	0	0.00%
- underexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 97. Overexplicitness with four (+) active referents

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

Unit: 3rd_person_subjects + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 + Set 7: natives +

	english1		greek1		english2		greek2		english3		greek3		natives	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PAS_IN_DISCOURSE	N=321		N=202		N=279		N=246		N=423		N=213		N=366	
- pas_notapp	311	96.88%	200	99.01%	275	98.57%	244	99.19%	413	97.64%	206	96.71%	324	88.52%
- intersentential	5	1.56%	2	0.99%	3	1.08%	1	0.41%	5	1.18%	5	2.35%	3	0.82%
- intrasentential	5	1.56%	0	0.00%	1	0.36%	1	0.41%	5	1.18%	2	0.94%	4	1.09%
INTRASENTENTIAL-TYP	N=321		N=202		N=279		N=246		N=423		N=213		N=366	
- main_sub	4	1.25%	0	0.00%	1	0.36%	1	0.41%	5	1.18%	2	0.94%	4	1.09%
- sub_main	1	0.31%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Figure 98. PAS cases

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

Unit: 3rd_person_subjects + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 + Set 7: natives +

	english1		greek1		english2		greek2		english3		greek3		natives	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PRAGMATICALITY	N=174		N=170		N=200		N=216		N=338		N=198		N=342	
- pragmatic	169	97.13%	168	98.82%	195	97.50%	212	98.15%	323	95.56%	190	95.96%	313	91.52%
- unpragmatic	5	2.87%	2	1.18%	5	2.50%	4	1.85%	15	4.44%	8	4.04%	29	8.48%
UNPRAGMATIC-TYPE	N=5		N=2		N=5		N=4		N=15		N=8		N=29	
- overexplicit	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
- underexplicit	5	100.00%	2	100.00%	5	100.00%	4	100.00%	15	100.00%	8	100.00%	29	100.00%
OVEREXPLICIT-TYPE	N=0		N=0		N=0		N=0		N=0		N=0		N=0	
- overt_when_null	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
- np_when_null	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
- np_when_overt	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
UNDEREXPLICIT-TYPE	N=5		N=2		N=5		N=4		N=15		N=8		N=29	
- null_when_overt	1	20.00%	0	0.00%	1	20.00%	1	25.00%	5	33.33%	1	12.50%	11	37.93%
- null_when_np	1	20.00%	2	100.00%	2	40.00%	2	50.00%	7	46.67%	4	50.00%	15	51.72%
- overt_when_np	3	60.00%	0	0.00%	1	20.00%	0	0.00%	2	13.33%	2	25.00%	2	6.90%
- np_ambiguous	0	0.00%	0	0.00%	1	20.00%	1	25.00%	1	6.67%	1	12.50%	1	3.45%

Figure 99. Underexplicitness by group

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

Unit: pragmatic + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 + Set 7: natives +

	english1		greek1		english2		greek2		english3		greek3		natives	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PROTAGONIST	N=169		N=168		N=195		N=212		N=323		N=190		N=313	
- protagonist_no	87	51.48%	23	13.69%	88	45.13%	61	28.77%	206	63.78%	122	64.21%	159	50.80%
- protagonist_yes	82	48.52%	145	86.31%	107	54.87%	151	71.23%	117	36.22%	68	35.79%	154	49.20%
SHARED_KNOWLEDGE_C	N=169		N=168		N=195		N=212		N=323		N=190		N=313	
- shared_no	154	91.12%	163	97.02%	188	96.41%	203	95.75%	288	89.16%	165	86.84%	272	86.90%
- shared_yes	15	8.88%	5	2.98%	7	3.59%	9	4.25%	35	10.84%	25	13.16%	41	13.10%

Figure 100. Protagonist and shared knowledge in pragmatic subjects

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

Unit: underexplicit + Show

Set 1: english1 + Set 2: greek1 + Set 3: english2 + Set 4: greek2 + Set 5: english3 + Set 6: greek3 + Set 7: natives +

	english1		greek1		english2		greek2		english3		greek3		natives	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
PROTAGONIST	N=5		N=2		N=5		N=4		N=15		N=8		N=29	
- protagonist_no	1	20.00%	0	0.00%	4	80.00%	1	25.00%	9	60.00%	4	50.00%	14	48.28%
- protagonist_yes	4	80.00%	2	100.00%	1	20.00%	3	75.00%	6	40.00%	4	50.00%	15	51.72%
SHARED_KNOWLEDGE_C	N=5		N=2		N=5		N=4		N=15		N=8		N=29	
- shared_no	4	80.00%	2	100.00%	4	80.00%	2	50.00%	5	33.33%	2	25.00%	4	13.79%
- shared_yes	1	20.00%	0	0.00%	1	20.00%	2	50.00%	10	66.67%	6	75.00%	25	86.21%

Figure 101. Protagonist and shared knowledge in underexplicit subjects

RESUMEN

Es un hecho ampliamente aceptado que el uso y la alternancia de las formas anafóricas en el discurso (pronombres nulos, pronombres plenos, sustantivos, etc.) están sintáctica y contextualmente condicionados. También se ha demostrado que los aprendices adultos de varios idiomas muestran déficits en cuanto a la interpretación y distribución de sujetos anafóricos. Mientras que la investigación en este campo se ha centrado tradicionalmente en la resolución/interpretación anafórica (en contraposición a la producción) y la mayor parte de los resultados se basan en datos experimentales, hay un número considerable de investigadores que señalan la necesidad de utilizar corpus electrónicos de aprendices para comprobar las hipótesis existentes de adquisición de segundas lenguas. Además, la mayoría de los estudios previos de anáfora en español L2 han examinado la interlengua de estudiantes ingleses, cuya L1 difiere del español con respecto a la gama de expresiones referenciales (el inglés, al contrario que el español, es una lengua de sujeto pleno). Por otro lado, los pocos estudios que se centran en estudiantes no anglófonos se ocupan generalmente solo de la interpretación de sujetos anafóricos. En general, en la adquisición del español L2, hay un número muy limitado de estudios orientados a la producción y centrados en la interlengua de aprendices con lengua materna de sujeto nulo (como es el griego, el árabe o el italiano).

Esta tesis tiene como objetivo explorar el uso anafórico de la 3ª persona en la interlengua de estudiantes ingleses y griegos de español L2 en varios niveles de competencia. Además, este estudio tiene como objetivo proporcionar una explicación general sobre los factores que condicionan las elecciones referenciales en español L1. El enfoque teórico integrado que se ha adoptado aquí se basa en propuestas relevantes de lingüística teórica, psicolingüística, lingüística computacional y lingüística de corpus. La base de datos empírica de esta investigación es CEDEL2, un corpus electrónico que contiene datos de producción escrita de estudiantes de español L2 de origen inglés y griego. Además, CEDEL2 contiene datos de hablantes nativos de español como un corpus de control. Crucialmente, los tres componentes de CEDEL2 exhiben los mismos principios de diseño. Por lo tanto, este es el primer estudio basado en un corpus de español L2 que compara tres niveles de competencia de dos grupos de alumnos (cuyas lenguas maternas difieren con respecto a la distribución de sujetos anafóricos) frente a un grupo de control. El objetivo principal de esta tesis es examinar varias hipótesis de adquisición de segundas

lenguas, centrándose en el papel de la transferencia de la lengua materna en la anáfora discursiva.

La herramienta UAM CorpusTool fue utilizada para la anotación y análisis de los datos del corpus. Con este propósito, se diseñó un conjunto de etiquetas y, a continuación, se realizó una anotación detallada sobre aspectos lingüísticos específicos, siguiendo la metodología utilizada en estudios previos basados en corpus electrónicos de aprendices. Se examinaron aprendices de tres niveles diferentes de competencia (intermedio, avanzado y muy avanzado) de cada origen (inglés y griego) y se compararon con el grupo de control. Los resultados han demostrado que, aunque los estudiantes de origen griego de nivel intermedio y avanzado muestran una cierta tendencia a la sobreutilización de sujetos anafóricos, lo hacen en un porcentaje significativamente menor que sus homólogos ingleses. Por otra parte, en el nivel muy avanzado, los estudiantes de origen griego exhiben preferencias nativas, en contraste con los aprendices de origen inglés, que son redundantes incluso en los niveles más altos de competencia linguistica. La influencia de la lengua materna puede explicar estas diferencias entre los dos grupos de estudiantes. Los estudiantes de origen griego parecen aprovechar de la similitud entre su L1 y el español con respecto a la distribución de sujetos anafóricos, mientras que los estudiantes de habla inglesa parecen transferir las propiedades correspondientes de su L1. Sin embargo, el hecho de que los estudiantes de nivel intermedio de habla griega también estén ocasionalmente redundantes está en línea con dos postulaciones que han sido presentadas muy recientemente en la literatura de adquisición de segundas lenguas. En primer lugar, la redundancia puede ser una tendencia universal en los niveles intermedios de competencia. En segundo lugar, ningún factor único puede explicar con éxito las deficiencias de los alumnos y sólo la consideración de múltiples factores que actúan a la vez puede dar plenamente cuenta del rendimiento observado.

CONCLUSIONES

Las conclusiones alcanzadas y la contribución general del presente estudio a la investigación actual sobre anáfora se resumirán en esta sección. Además, se señalarán algunas limitaciones de este estudio y se sugerirán algunas direcciones para investigaciones futuras.

Resumen de conclusiones

El presente estudio se ha centrado en el fenómeno lingüístico de la anáfora en español L1 y L2. Se ha hecho especial hincapié en la adquisición de sujetos anafóricos de 3ª persona por aprendices adultos de español L2. Para este propósito, este estudio es el primero en examinar grupos de aprendices de dos orígenes diferentes (inglés y griego) en tres niveles de competencia lingüística (intermedio, avanzado y muy avanzado). Además, se ha examinado la relevancia de varios factores sintácticos y discursivos en la producción de formas anafóricas en español L1. La metodología de corpus electrónicos empleada en este estudio nos ha permitido examinar una serie de afirmaciones e hipótesis hechas previamente en la literatura con respecto a la anáfora discursiva y la adquisición de segundas lenguas. En este capítulo presentamos las conclusiones alcanzadas y sus implicaciones para la investigación relevante. Además, se discuten algunas de las limitaciones del presente estudio y se sugieren algunas direcciones para futuras investigaciones.

Las conclusiones alcanzadas en el presente estudio, en relación con las hipótesis de investigación, se pueden resumir de esta forma:

I. La producción de sujetos anafóricos de 3ª persona en español L1 puede ser debidamente explicada como resultado de la compleja interacción entre múltiples factores sintácticos y discursivos.
II. Los estudiantes de español L2 muestras déficits con algunas propiedades implicadas en la producción de sujetos anafóricos de 3ª persona. Estos déficits pueden ser superados por algunos estudiantes.
III. Los principales déficits detectados en la producción anafórica de los estudiantes de español L2 conciernen la producción de sujetos redundantes de 3ª persona.

IV. El nivel de competencia afecta de manera decisiva al rendimiento de los estudiantes de español L2, de forma que los grupos más competentes obtienen mejores resultados que los grupos menos competentes.

V. La L1 es un factor crucial en la adquisición de sujetos anafóricos de 3ª persona.

VI. El desempeño no nativo de los estudiantes de español L2 puede ser mejor explicado en términos de la influencia de múltiples factores.

Limitaciones e investigaciones futuras

Es crucial señalar también algunas limitaciones del presente estudio que podrían superarse en futuras investigaciones:

- Primero, este estudio ha examinado algunas combinaciones lingüísticas particulares con un número limitado de participantes en un género y modo de discurso específico. Nuestros hallazgos deben ser comprobados en el futuro con otras combinaciones de idiomas (por ejemplo, estudiantes de italiano L2 de origen griego). Estudiantes con menor nivel de competencia (por ejemplo, principiantes) deben ser también incluidos en el grupo de participantes en estudios futuros y se debe examinar a más participantes por grupo. Deberían examinarse también diferentes géneros y modos de discurso (por ejemplo, conversaciones orales).
- En segundo lugar, como han señalado recientemente algunos autores, las investigaciones orientadas a la producción siempre deben complementarse con estudios experimentales. Recordamos que los resultados del presente estudio se refieren exclusivamente a la producción de sujetos anafóricos en discurso real. Por lo tanto, no se pueden hacer reclamaciones con respecto a la interpretación de sujetos anafóricos y el trabajo experimental en esta dirección es necesario para triangular hallazgos y proporcionar una cuenta completa de la anáfora en la adquisición de primera y segunda lengua. Sobre la base de los hallazgos del presente estudio podrían diseñarse experimentos en el futuro con el fin de probar cada una de las propiedades particulares examinadas aquí (por ejemplo, la resolución de sujetos anafóricos según el tipo de cláusula, el número de referentes, etc.)
- En tercer lugar, con respecto al análisis de los datos de español L1, el enfoque metodológico adoptado aquí no nos has permitido examinar ni la interacción de los diferentes factores entre sí ni las diferencias potenciales de entre todas las

formas de sujeto pleno. La investigación futura sobre anáfora discursiva debería beneficiarse de la aplicación de modelos estadísticos sofisticados en corpus electrónicos grandes, en estrecha cooperación con expertos de otros campos disciplinarios (por ejemplo, de estadística), con el fin de dar cuenta de la anáfora en toda su complejidad.

- En cuarto lugar, el papel de lenguas extranjeras previamente aprendidas (influencia de la L2 a la L3) no fue específicamente abordado en el presente estudio. La investigación futura debe determinar hasta qué punto este factor afecta el rendimiento de los aprendices de segundas lenguas. Además, debe determinarse cómo funcionan los factores interactivos que afectan la adquisición de los sujetos anafóricos (así como la adquisición de cualquier otro fenómeno lingüístico) por diferentes pares de lenguas fuente/objetivo y niveles de competencia.
- Finalmente, aunque no se haya controlado en el presente estudio, también pueden estar presentes diferencias individuales en la producción de sujetos anafóricos (así como en la adquisición de cualquier otro fenómeno lingüístico). En el futuro, la investigación cualitativa debe determinar hasta qué punto las preferencias individuales (por ejemplo, el estilo) pueden afectar las opciones referenciales de hablantes nativos y estudiantes de lenguas extranjeras.