PRONOMINAL SUBJECTS
IN PENINSULAR SPANISH
AND EUROPEAN PORTUGUESE

PEKKA POSIO
Pronominal subjects in Peninsular Spanish and European Portuguese

Semantics, pragmatics, and formulaic sequences

Pekka Posio

ACADEMIC DISSERTATION

To be publicly discussed, by due permission of the Faculty of Arts at the University of Helsinki in Arppeanum (Auditorium, Snellmaninkatu 3), on the 2\textsuperscript{nd} of November 2012 at 12 o’clock
Abstract

The present dissertation examines the expression of subject personal pronouns in Peninsular Spanish (PS) and European Portuguese (EP). The goals of the research are (1) to provide empirical data from PS and EP and compare them, (2) to present a theoretical account of the findings and (3) to deepen understanding of factors affecting subject pronoun use in languages with variable subject expression. The theoretical framework is cognitive-functional and usage-based linguistics: grammar is understood as emerging through recurrent practices used by the speakers and grammatical variation is assumed to be shaped by cognitive and interactional properties of language use.

The dissertation consists of four articles, examining data from different corpora of spoken PS and EP, and an introductory chapter. The research is motivated by the findings of previous studies suggesting that there is a significant relation between variable subject expression and verb lexemes or semantic roles of the subject. The methodology combines qualitative examination of the data with quantitative frequency analyses. The main outcome of the research can be summarized as follows:

- PS and EP differ significantly with regard to the frequency of expressed subject pronouns. In both languages, the factors affecting the expression of pronominal subjects vary from one grammatical person to another.
- Subject omission does not constitute a default option in either of the languages. Rather, both expression and omission may serve different pragmatic functions. In addition, subject expression has acquired a formulaic character with frequently occurring verb tokens.
- In first person singular, subject omission is favored by agentivity of the subject whereas stative subjects are more often omitted, reflecting the focusing of attention on the subject or on other clausal participants.
- In first person plural, subject expression is rare in PS and typically occurs with reference to hearer-exclusive groups. Such restrictions are not found in EP.
- Postverbal placement of subject pronouns serves both contrasting and backgrounding functions. In EP, postverbal placement of pronominal subjects is mostly restricted to certain non-productive, formulaic sequences.
- In both languages, subject expression with the most frequently occurring verb tokens diverges from the general patterns of subject pronoun expression. In these cases, subject expression or omission has grammaticalized as part of the construction where it occurs.

In addition to describing subject expression in PS and EP, the present research contributes to the discussion of the expression of pronominal subjects in a wider typological and theoretical context. In previous research, null subject languages have been assumed to form a relatively homogeneous group where subject omission is the norm and expression is reserved e.g. for the expression of contrast and emphasis. The present research shows that while the factors affecting subject expression may be similar, even closely related languages differ significantly with regard to the rate of expressed vs. omitted subjects. It also suggests that formulaic, grammaticalizing constructions play an important role in accounting for subject expression in speech.
Acknowledgments

I have unfortunately forgotten the name of the person who provided the first incentive for writing this dissertation. She was a student attending my Portuguese course, about six or seven years ago, who had learned Spanish before Portuguese. “Isn’t it so that subject pronouns are used more often in Portuguese than in Spanish?” she asked. All I could answer her at that time was that I had the same feeling, too, but I did not know if anyone had ever investigated the matter. The question – simple as it seemed – started to intrigue me to the extent that it provided the topic for my master’s thesis. It is also one of the questions the present dissertation is trying to answer.

During the four years I have been working on this project I have had a privileged position as an early stage researcher, first at the Department of General Linguistics, now integrated into the Department of Modern Languages. I have also benefited from the proximity of teachers, colleagues and friends at the former Department of Romance Languages where my academic studies started over ten years ago. I owe my sincere gratitude to many people who have contributed to this dissertation project. I will try to mention as many of you as possible, although providing a comprehensive list would require more space than is available here.

First and foremost, I would like to express my gratitude to my supervisors Seppo Kittilä, who was my main supervisor during the whole PhD project, my second supervisor María José Serrano, and Angela Bartens who supervised the early stages of my work. This dissertation would never have been concluded without your continuous feedback and support. Your professionalism, productivity and enthusiasm have been a source of inspiration to me. The two preliminary examiners of the thesis, Catherine Travis and Richard Cameron, also provided invaluable feedback and helped to improve the scientific quality of the introductory chapter. Thank you for your inspiring comments and suggestions for revision, as well as your recommendations for future research.

I am grateful to all my officemates and colleagues at the General Linguistics’ corridor and beyond. It has been a pleasure to work with you. In particular, I would like to thank Fred Karlsson, who has been not only a great boss but also an excellent teacher and an inspiring colleague. Special thanks go also to Kaius Sinnemäki, who always found the time and patience to help me with several methodological problems I faced in my research. I would also like to express my gratitude to Maria Vilkuna for enjoyable cooperation and conversations on various topics including but not limited to linguistics.

I am grateful to Kone Foundation for providing me a research grant for the first year of my project, and to the LANGNET Doctoral Program in Language Studies. Being part of the program has provided me not only the necessary funding for my work but also the possibility to participate in enriching courses and seminars and, most importantly, get acquainted with and become a part of the community of people doing linguistic research in Finland. I would like to thank all my colleagues at LANGNET for making these four years a wonderful experience, among them especially Anton Granvik, Hanna Lantto, Aki Kyröläinen, Lotta Jalava and Max Wahlström. Special thanks go to Erika Sandman and Katja Västí with whom I have shared many conference trips, a supervisor, and a flat in Leipzig.
One of the best experiences during my dissertation project was the opportunity to spend three months as a guest researcher at the graduate college “Frequency effects in language” at the Albert-Ludwigs-Universität Freiburg in Germany. I would like to express my gratitude to the research community in Freiburg for their scientific and material support and, most importantly, for your interest in my work and for many fruitful and inspiring discussions we had and will continue to have. I am very grateful to many people in Freiburg, among them especially Aria Adli, Philipp Dankel, Daniel Jacob, Nikolay Khakimov, Stefan Pfänder and Stefan Savić. Thank you for making my stay possible and turning it into the valuable experience it was, both scientifically and personally.

Last but not least, I want to express my gratitude to all my friends and relatives, for offering their support and having the patience to listen to my worries related to various aspects of subject pronoun expression and/or academic life in general. Ida and Aino Koivisto, I will always miss our lunch and coffee time discussions! And of course I am eternally grateful to Maxim: thank you for sharing your life with me during this dissertation project and during the years to come.

Dear student whose name I have forgotten: if you happen to read this dissertation, I hope you find at least a partial and tentative answer to your wise question.

Helsinki, September 10\textsuperscript{th} 2012
Contents

Abstract ........................................................................................................................ i
Acknowledgments ....................................................................................................... ii
List of articles .............................................................................................................. v
1. Introduction ......................................................................................................... 1
   1.1. Objectives and initial expectations ................................................................. 3
2. Expression of pronominal subjects: some preliminaries ................................... 6
   2.1. Terminological issues .................................................................................... 6
   2.2. Subject expression in languages of the world ................................................. 8
3. Theoretical background .................................................................................... 11
   3.1. Research on Spanish .................................................................................... 11
      3.1.1. Contrast and emphasis ............................................................................. 13
      3.1.2. Grammatical persons and syncretisms ...................................................... 15
      3.1.3. Sociolinguistic and areal variation ........................................................... 17
      3.1.4. Information structure and switch reference ............................................... 18
      3.1.5. Semantics and pragmatics of subject expression........................................ 19
      3.1.6. Semantic and pragmatic salience or focusing of attention .......................... 21
      3.1.7. Priming effects ......................................................................................... 23
      3.1.8. Frequency effects ..................................................................................... 24
   3.2. Research on Portuguese ............................................................................... 26
      3.2.1. Expression of pronominal subjects in Brazilian and European Portuguese 27
      3.2.2. Differences between Brazilian and European Portuguese and Spanish ...... 31
4. Data and methodology ...................................................................................... 33
   4.1. Typological profile of Peninsular Spanish and European Portuguese .......... 33
   4.2. The data and their comparability .................................................................. 36
   4.3. Methods ...................................................................................................... 38
5. Results and conclusions ..................................................................................... 40
   5.1. Summary of the articles ............................................................................... 40
   5.2. Peninsular Spanish and European Portuguese in comparison .......... 42
   5.3. Subject expression, verbs and grammatical persons ..................................... 44
   5.4. Subject pronoun expression, semantics and pragmatics............................... 47
   5.5. Conclusions ................................................................................................. 48
References ................................................................................................................. 50
Corpora ..................................................................................................................... 55
List of articles

I

II

III

IV
1. Introduction

It has been claimed that “[s]ubject expression is one of the most widely studied features of Spanish syntax, yet it remains one of the least understood” (Travis 2007:106). Indeed, the variable use of pronominal subjects in so-called null subject languages, especially Spanish, has intrigued scholars of different theoretical and empirical orientations in recent decades and there is a vast array of studies on subject expression focusing on different formal or functional aspects of the phenomenon. In brief, the problem of subject expression can be summarized through the following questions:

1. When and why do speakers express pronominal subjects or leave them unexpressed?
2. If pronominal subjects are expressed, what conditions their position before or after the verb?

These questions are illustrated by example (1) where all clauses contain the same verb, creer ‘think, believe’ in first person singular. As can be observed, the subject is expressed in the preverbal position in (1a), in the postverbal position in (1b), and left unexpressed in (1c).

(1) a. Bueno, yo creo que ante todo... hay que separar
   well I think-1SG COMPL before all have-3SG COMPL separate-INF
   ‘Well, I think that to begin, you have to separate
   la vida privada con la vida profesional...
   the life private with the life professional
   your personal life from your professional life.’
   COREC: EENT005L

b. ... pues aquí se trabaja muy bien, creo yo.
   well here RFL.3 work-3SG very well think-1SG I
   ‘well, people work very well here I think.’
   COREC: PENT001D

c. – Es que... perdona, ¿tú tienes hijos?
   be-3SG COMPL sorry you have-2SG children
   ‘It’s that… sorry, do you have children?’

   – Tengo... no los he contado, pero creo que son... cinco.
   have-1SG NEG ACC.3PL have-1SG counted but think-1SG COMPL be-3PL five
   ‘I have... I haven’t counted them, but (I) think there are five of them.’ (COREC: ENT007D)

Despite the extensive literature on subject pronoun expression it is not clear which factors play the most crucial role in accounting for the use or omission of subject pronouns. Although same/switch reference has been found to affect subject expression in all dialects and all persons in all studies on Spanish where it has been one of the explanatory factors (see e.g. Silva-Corvalán 1982, 1997, 2001; Enríquez 1984; Hochberg 1986; Morales 1989; Bentivoglio 1987; Cameron 1992, 1993, 1995, 1997; Flores-Ferrán 2002, 2004; Travis 2005, 2007), it does not exhaust the variation found in subject expression as there are many other factors that have been shown or suspected to affect the phenomenon.
The present study does not claim to solve all the problems that have arisen in previous research, nor does it aspire to study all factors possibly affecting subject pronoun usage in order to draw a general picture of the phenomenon. Rather, the scope of observation is limited in order to focus on a question that has received relatively little attention in the previous studies although it has proved to have a significant effect on the usage of pronominal subjects: what verb lexemes or forms tend to occur frequently either with or without pronominal subjects, and why. As for grammatical persons, the main focus is on first person singular where previous research has pointed at a systematic connection between verb semantics and subject pronoun expression (cf. Cameron 1992: 48). However, along with first person singular, the articles also examine second person singular (Posio 2011), first person plural (Posio to appear) and other grammatical persons (Posio 2012a). The research questions are related to semantic and pragmatic properties of different verb lexemes and the main interest lies in the ways speakers actually use first person singular verbs and pronouns in discourse. It is shown that pragmatic, formulaic uses of frequently occurring verbs are crucial in accounting for the use of subject pronouns in first person singular. The present study does not treat questions related to anaphora or information structure, although their importance with regard to the differences between deictic and anaphoric persons is discussed (see Section 5.3).

Another important aspect of the current study is the comparison between two null subject languages, Peninsular Spanish and European Portuguese. If subject pronoun usage in Spanish is still poorly understood despite extensive research on the phenomenon (as claimed in Travis 2007: 106), the situation is much worse with regard to Portuguese, especially the European variety. Subject pronoun usage in European Portuguese has been studied very little, and in most cases the objective has been to compare Brazilian and European Portuguese (e.g. Duarte 1995; Barbosa, Duarte & Kato 2001, 2005) in the interest of demonstrating that Brazilian Portuguese is gradually losing its null subject properties while European Portuguese remains a prototypical null subject language. However, there are no previous corpus-based empirical studies concentrating solely on European Portuguese or on the comparison between European Portuguese and Spanish. It should also be noted that most of the previous research on subject pronoun usage in Spanish has concentrated on American Spanish and less attention has been paid to the Peninsular variety.

Given the scarcity of empirical studies, relatively little is known of the expression of pronominal subjects in European Portuguese, especially with regard to the eventual connection of subject expression with different verb lexemes or forms. It is thus crucial to carry out empirical research using authentic data from spoken language in order to find out to what extent subject pronoun usage is similar in European Portuguese and Peninsular Spanish. Both of the languages under survey are regarded as null subject languages, more precisely put as languages where subject person is expressed by verbal affixes and the use of independent pronominal subjects is optional (Dryer 2011). However, as the present study shows, in spite of being closely related, Peninsular Spanish and European Portuguese differ significantly with regard to the frequency of subject expression and, to a certain extent, with regard to the semantic and pragmatic factors affecting subject pronoun usage.
This dissertation adheres to the principles of cognitive-functional linguistics and the usage-based view of language. Thus grammar is understood as emerging through recurrent practices used by the speakers and grammatical variation is assumed to be functionally motivated and shaped by cognitive and interactional properties of language use (e.g. Hopper 1987, 1998, 2011; Bybee 2010, Croft & Cruse 2004; Travis & Torres Cacoullos 2012). Grammar is not regarded as a rule-based system that generates the possible sentences of a language, but rather as “observed repetitions in discourse” (Hopper 1998:156), a series of conventionalized usage patterns that become entrenched to various degrees through repetition. In order to answer the study questions, the research aims at finding out what speakers actually do with first person singular verb forms in and how the presence or absence of the subject pronoun relates to the communicative functions of these forms. On the other hand, while speakers use linguistic elements for different functional purposes, they are also affected by their experience of the usage by other speakers. This is manifested in the form of different frequency effects (Bybee & Thompson 1997, Bybee 2010).

In line with the usage-based approach, the present work examines emerging phenomena\(^1\) such as the development of formulaic sequences (especially in Posio 2012a and to appear) and the effects of frequency (Posio to appear). The methods used in the study are primarily qualitative, but quantitative data is extracted and analyzed using appropriate statistical tests (Posio 2011, 2012b, to appear). Combining quantitative and qualitative approaches is crucial in order to understand the tendencies observed in the data. In all articles included in this dissertation, quantitative examination of the data is used to identify frequently occurring patterns. However, as explanations for the phenomena identified are situated at the pragmatic and interactional level, careful qualitative scrutiny of examples in their context is needed in order to provide an account of the usage patterns identified.

1.1. Objectives and initial expectations

The principal goal of this dissertation is to provide new information of the use of pronominal subjects in Peninsular Spanish and European Portuguese. In the case of European Portuguese, little descriptive work on subject expression has been done so far. Thus an important contribution of this study is to provide descriptive analysis of European Portuguese subject expression. A comparison with another null subject language such as Peninsular Spanish provides a fruitful starting point (see Posio 2012a, 2012b, to appear). While the expression of pronominal subjects has been studied extensively in different (mostly Latin American) varieties of Spanish, less attention has been paid to the different usage patterns with different verb lexemes that are given prominence in this work. Semantic class of the verb has been shown to have a significant effect on the expression of pronominal subjects in various studies (see Section 2.3) but in previous research this particular factor has not been studied

---

\(^1\) Hopper (2011) establishes a distinction between the terms “emerging” and “emergent”, using “emerging” to refer to the process whereby recognizable grammatical structures are established through usage, whereas “emergent” is connected to the process in which “speakers arrange constructional patterns in a novel and improvised way” (Auer & Pfänder 2011:5) in different communicative situations.
extensively enough in order to provide an explanation for the observed influence of the verb class. The comparison with European Portuguese serves to highlight the special properties of subject expression in Spanish that may have remained unnoticed before (see especially Posio 2012b).

Apart from describing the two languages under survey, the work also contributes to the typological description of null subject languages in general. Although the perspective adopted is not strictly typological, the comparison of data drawn from two genealogically, typologically and geographically close language varieties calls into question the assumption that null subject languages form a uniform class. Rather, the findings of the present work suggest that although the factors affecting subject expression in the two languages under survey are to a certain extent similar, there are also language-specific tendencies that can only be revealed by examining a sufficient amount of data drawn from actual language use (Posio 2011, 2012a, 2012b, to appear). Instead of just assuming that closely related languages behave similarly, it is worth observing these languages using authentic data. From a methodological perspective, the current work also tests and discusses the adaptability of semantic roles to the analysis of authentic data from spoken language (see Posio 2012a).

Initially, prior research reviewed in Posio (2008, 2010) provides the following expectations:

1. PS and EP differ systematically from each other with respect to the frequency of subject expression in 1SG (Posio 2008, 2010).

2. In both PS and EP, there is a connection between verb semantics and subject expression (Posio 2008, 2010). This difference can be operationalized through verb lexemes or, for the purposes of cross-linguistic comparison, semantic roles of the subject. These differences may be due to (a) different perceptual salience of the clausal participants, (b) different pragmatic uses of semantic roles or (c) different pragmatic uses of verb lexemes.

3. In both PS and EP, subject expression tends to become fixed in formulaic sequences with certain verb lexemes or verb forms. The development of such formulaic sequences is sensitive to frequency effects: the most frequent verb lexemes or forms present special tendencies of subject expression. The development of formulaic sequences may consequently lead to grammaticalization of patterns consisting of a verb form either with or without an expressed pronominal subject (cf. Travis 2006).

4. The factors affecting the expression of pronominal subjects are different in different grammatical persons. For instance, in third person null subjects are allowed if the subject referent is topical or accessible because it has been previously introduced into the discourse. In first person singular, the referent is always accessible by virtue of being present in the communicative situation (Ariel 2001), and need not be introduced into the discourse. However, previous research

---

2 The four articles included in the dissertation continue the research started in my master’s thesis (Posio 2008; part of the results were published in Posio 2010) in which various factors potentially affecting the use of first person singular subject pronouns in Peninsular Spanish and European Portuguese were examined. The study pointed at some interesting differences in the frequency of subject pronoun expression between the languages but did not provide an explanation for them.
indicates that subject expression tends to be more frequent in first person singular rather than in the other persons (see Table 2 on page 15). This apparent mismatch between the inherent level of accessibility and frequency of subject expression suggests that accessibility may not be the key factor in accounting for subject expression at least in the case of first person singular.

A crucial question with regard to all of the hypotheses listed here is the comparability of the data: this issue is addressed explicitly in Section 4.2.
2. Expression of pronominal subjects:
   some preliminaries

This section discusses the terminology related to expression of pronominal subjects. In the articles of the present dissertation, the term null subject language is used to refer to such languages where finite clauses may lack overtly expressed, independent subject morphemes. A better terminological alternative would be to speak of languages with variable subject expression, as subject pronoun expression need not be in any way more marked or less frequent than their omission in a language pertaining to this category, and it is indeed subject to considerable variation.³

Null subject languages are not a typologically uniform class with regard to the coding of the subject argument or the frequency of subject expression. Defining null subject languages is thus not a trivial task. Section 2.1 discusses some terminological choices in the current work and Section 2.2 examines the types of pronominal subject expression found in the languages of the world, in order to situate Spanish and Portuguese into a wider typological context.

2.1. Terminological issues

Subject expression is a widely studied phenomenon and various terms have been used to refer to it in previous scholarship. Although the choice of terminology used to speak of the phenomenon is indicative of the theoretical framework adopted, some terms such as “pro-drop” or “null subject” are widely used within different approaches. The widespread use of the term “pro-drop” originates in the Principles and Parameters version of the generative framework where it was used to speak of a cluster of properties including the possibility of null subjects and “free” subject-verb inversion (Chomsky 1981). The term is nevertheless commonly used outside this framework to refer to only one property of the parameter, i.e. null subjects. In addition to subjects, it may also refer to the “dropping” of other arguments than subjects: East Asian languages such as Chinese and Japanese are oft-cited examples of languages commonly allowing the absence subjects and objects when their referents are identifiable by the context. However, in its most common usage, “pro-drop” is understood as referring to the lack of subjects expressed by independent morphemes (i.e. expressed only by affixes or clitics or not overtly expressed at all).

A further distinction can be made between referential and expletive null subjects: in generative accounts, impersonal clauses such as Spanish llueve ‘it rains’ may be analyzed as containing a null expletive subject that never surfaces as an overt form.⁴ It is not always clear whether these kinds of clauses necessarily lacking independent subjects should be analyzed as “pro-drop” or not.

³ I would like to thank Catherine Travis (p.c.) for suggesting me the use of this term in future research.
⁴ As for Portuguese, although standard EP does not manifest overt expletive subjects, in dialectal EP the masculine third person singular subject pronoun ele may be used expletively in impersonal constructions (cf. Carrilho 2005).
The term pro-drop is not used in the current work. Apart from evoking a processual metaphor considered inadequate in the current perspective, the term is associated with a different theoretical background and it is unclear whether its use implies the possibility of “dropping” non-subject arguments and other properties attributed to the pro-drop paradigm. The preferred terms used in this dissertation are subject pronoun usage and presence or absence of pronominal subjects. This terminological choice highlights the idea that speakers actually use the presence or absence of the subject pronoun for some purpose and neither the use nor the omission should be considered a priori as a default option. This being said, it has to be admitted that it is often difficult to speak of the lack of something without referring to its presence. Such expressions as “omitting the subject” and “null” or “overt” subject are difficult to avoid completely, and they do occur in all of the articles included in the dissertation. In the current context, such expressions should nevertheless be understood metaphorically: they are not intended to suggest that there is an underlying subject that is “omitted” or that there is a “null pronoun” lacking phonological manifestation in the absence of overt marking.

Languages differ considerably with respect to the morphosyntactic marking of pronominal subjects. In some null subject languages such as Chinese and Japanese there is no person marking morphology apart from independent pronominal subject words. In other languages, such as Portuguese and Spanish, person marking morphology on the verb most often permits the identification of the subject referent even when there is no independent subject in the clause (see Section 2.2).

Person marking morphology may be analyzed as agreement, i.e. the verb is said to agree with its subject, or as bound person marking: both terms are acceptable and the choice depends on the purpose of the analysis and the theoretical perspective adopted. In languages where independent subjects need not be expressed, the latter term seems more appropriate if the presence of a null pronominal element (that the verb could agree with) is not postulated. Some scholars thus identify bound person markers as subjects and independent subject pronouns as facultative adjuncts (Van Valin & LaPolla 1997: 331-332). However, bound person markers do not share many of the syntactic and pragmatic properties of prototypical subjects. They cannot appear independently of the verb or bear prosodic stress, occur in the preverbal position, control reflexive and possessive pronouns, or be omitted. Dixon (2010:39), among others, uses the terms “free pronouns” and “bound pronouns”: if both are present in a clause, the clause is said to have discontinuous subject marking. In the present work, independent subject pronouns are considered syntactically as subjects whereas person marking affixes are analyzed as bound person markers.

---

5 Note also that the terms “subject pronoun” and “pronominal subject” are used interchangeably in this work and, unless otherwise indicated, they refer to subject personal pronouns (and not e.g. to demonstrative pronouns).
2.2. Subject expression in languages of the world

One of the early uses of the term null subject language is found in the generative literature, where it refers to languages where subject pronouns were supposed to be avoided in line with the “Avoid Pronoun Principle” (Chomsky 1981:65) when the referent of the subject is fully identifiable without them (Fernández Soriano 1999; Barbosa, Duarte & Kato 2005). Like “pro-drop”, the term null subject language is nevertheless used widely within other theoretical frameworks as well. From a typological perspective, the term is somewhat imprecise and may actually be used to refer to languages with highly divergent ways of subject marking. Languages requiring independent subject pronouns to be expressed are rather infrequent outside the branch of Germanic languages spoken in North-Western Europe. There is thus nothing exceptional or unusual with the fact that independent personal pronouns need not be expressed. Indeed, conceptualizing subject pronoun absence as “deletion” or “dropping” has been rightly criticized of linguistic Anglo-centrism, i.e. “analyzing other languages as being underlyingly like English despite their superficial differences” (Dryer 2011).

Dryer (2011), in his typological classification of a sample of 711 languages, does not use the term null subject language at all but rather distinguishes between six different strategies used in the expression of pronominal subjects. Spanish and Portuguese are included in the largest category, among languages “in which the normal expression of pronominal subjects is by means of affixes on the verb” (Dryer 2011). This group includes 437 of the 711 languages in Dryer’s sample (i.e. 62%). Dryer’s (2011) classification is shown in Chart 1.

Chart 1. Distribution of different strategies of subject person expression (Dryer 2011)

If null subject languages are understood to include languages where the “normal” way to express pronominal subjects is by affixes on verbs (N= 437) or by subject clitics on variable hosts (N= 32) and languages where subject pronouns are “often” left out and there is no marking by affixes or clitics (N= 61), in total 530 of the languages in the sample (i.e. 75%) are null subject languages. In a typological perspective, languages
requiring the presence of independent subject pronouns are thus a minority. Instead of trying to account for the conditions under which null subjects are licensed in null subject languages, it might actually be more relevant to ask why double marking is required in such typologically rare languages as English or German.

Dryer’s (2011) classification does not contemplate the fact that the use of independent pronominal subjects in languages where it is not obligatory is subject to a great deal of variation. If actual frequency of subject expression was considered, the classification of some languages might need to be reconsidered (see below). Although Dryer’s (2011) classification is not based on frequency counts but rather on reference grammars, frequency is mentioned at least implicitly at various points: subject pronouns are said to be “typically left out” or “normally if not obligatorily present”, null subjects are said to occur “seldom in actual usage” in some languages, and so on. It is also stated explicitly that “if all sentences with pronominal subjects on a couple of pages of text in a language have a pronoun in subject position, the language is coded as being of the first type [i.e. a language with obligatory subject pronouns]” (Dryer 2011).

Such formulations are troubling, as it is not said what material has been used for the frequency count: the only sources cited are reference grammars and other descriptive studies, not text or speech corpora. In order to establish a classification based on frequency of subject expression in actual language use, sufficiently comparable frequency data should be available from all languages included in the sample. In addition, frequency of subject expression is a problematic classification criterion if the purpose is to divide languages into discrete categories. Where exactly should one place the threshold? If, for instance, 50% of the pronominal subjects are expressed in a given amount of text, should the language be considered as a null subject language or not?

There are relatively few studies seriously examining the frequency of different subject expression strategies in a cross-linguistic perspective: a comprehensive typological study based on actual frequency data of the presence of pronominal subjects is yet to be done. This is understandable, as the availability of frequency data is mostly limited to large European languages and the comparability of data drawn from different languages is often questionable. Frequency of subject pronoun presence in different dialects of Spanish has been compared in several studies (see Flores-Ferran 2007 and references therein) and subject pronoun usage in European and Brazilian Portuguese has also been compared in previous research (Duarte 1995, 2000; Barbosa, Duarte & Kato 2005) but these studies focus on language-internal variation rather than on cross-linguistic differences. Spanish and Portuguese have been examined contrastively by Soares da Silva (2006) and Posio (2008, 2010, 2012a, 2012b, to appear).

Seo (2001) provides a comparative account of five Slavic languages, based on quantitative examination of parallel corpora consisting of literary texts translated from English. What these comparative studies suggest is that null subject languages diverge considerably with regard to the actual frequency of subject pronoun usage. For instance, Seo (2001) shows that the five languages examined differ significantly with regard to the frequency of the expression of pronominal subjects. Thus consider Table 1:

---

6 The frequencies do not include subjects judged “emphatic” because “an emphatic subject always tends to be overt no matter whether the language is a null subject or a non-null subject language” (Seo 2001: 70).
Table 1. Frequency of subject pronoun expression in Slavic languages (Seo 2001)

<table>
<thead>
<tr>
<th>Language</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polish</td>
<td>7%</td>
<td>93%</td>
</tr>
<tr>
<td>Czech</td>
<td>9%</td>
<td>81%</td>
</tr>
<tr>
<td>Serbo-Croatian</td>
<td>16%</td>
<td>84%</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>22%</td>
<td>78%</td>
</tr>
<tr>
<td>Russian</td>
<td>78%</td>
<td>22%</td>
</tr>
</tbody>
</table>

All languages in Seo’s (2001) sample except Russian are considered null subject languages. The South Slavic languages Bulgarian and Serbo-Croatian show a surprisingly high percentage of subject presence, while Polish and Czech have the highest percentages of null subjects. Interestingly enough, Russian, which is usually regarded as a non-null subject language, has a subject pronoun only in 78% of the clauses where a pronoun could be present. The considerable variation found leads Seo (2001: 94) to conclude that “the null-subject phenomenon is not an absolute value for human languages, but rather is a matter of the degree of the phenomenon in each language”. The term “null subject language” is best understood as a radial category with no clear-cut boundaries: this is the sense in which the term is used in the current work.

It is often assumed that languages are shaped by two competing motivations: a principle of economy or minimization and a principle of informativity or recipient design (see e.g. Hopper & Traugott 2003:71-73, Sacks & Schegloff 1979). In the case of subject expression, some languages are extremely “economic” and permit complete lack of subject expression, as is the case in Chinese and Japanese, while other languages like German and Icelandic are more on the informative side and require double marking of the subject.

Neither economy nor informativity seems to have great explanatory power in accounting for tendencies of subject pronoun usage within one language. It is often claimed, for instance, that subject pronouns must be left out when they are not needed for the identification of the referent (in accordance with the principle of economy or, more specifically, the “Avoid Pronoun Principle” of Chomsky 1981) or that they must be used when there is risk of ambiguity e.g. due to syncretic person marking. At least as far as Spanish and Portuguese are concerned, there is meager evidence to support such claims. Results such as those presented in Seo (2001) and in the present study suggest that null subject languages differ with regard to the actual rate of subject expression without there being any apparent functional motivation (such as syncretic person marking; see Section 3.1.2) behind these differences.

---

7 Polish differs from the others in that subject person is expressed by clitics attaching to variable hosts.
3. Theoretical background

This section presents a critical overview of the previous research and theories on subject pronoun usage in Spanish and Portuguese in order to situate the current work in the context where it belongs. Section 3.1 concentrates on Spanish and Section 3.2 on Portuguese. Given the existence of a considerable body of research conducted on the expression of pronominal subjects in Spanish, it is impossible to reference “all” previous studies here. Thus previous scholarship focusing on purely syntactic description of null subjects, with no consideration of actual usage, is mostly left out of the scope of this discussion. Another area where subject expression constitutes an important research topic but which nevertheless is omitted here is language acquisition. The perspective on previous research adopted in this section is not chronological: rather, the discussion is organized under headings referring to different explanatory factors suggested.

3.1. Research on Spanish

In most previous research, attention is focused on the expression of subject pronouns, not on their absence or, simply, their use. It is a basic assumption that it is the presence of subject pronouns that has to be explained in null subject languages, as subject expression is considered as the less frequent and more “marked” option. Dixon (2010:117) formulates the idea into a cross-linguistic generalization:

When bound pronouns are obligatory, the free forms [i.e. subject pronouns] are used sparingly, typically for emphasis or to mark the introduction of a new participant into a discourse.

It is nevertheless not clear whether subject expression should always be considered as rare or marked as opposed to omission because in some local contexts subject omission is actually less frequent than expression (e.g. with the verb creer in Spanish; cf. Travis 2006, Posio 2011). Rather than saying that one of the two options is “marked”, it is more fruitful to consider that both the expression and the omission of pronominal subjects may serve specific functions in discourse.

The descriptive grammar of the Real Academia Española (Fernández Soriano 1999:1227) gives the examples reproduced here as (2) and (3) to illustrate a use of subject pronouns judged as “totally impossible” in a null subject language like Peninsular Spanish:

(2) Juan es mi vecino de al lado. Él es estudiante de matemáticas.
*Juan is my next-door neighbor. He studies Mathematics

pero él se interesa también mucho por la filosofía porque
but he REF.3 interest-3SG also much PREP ART philosophy because

pero él se interesa también mucho por la filosofía porque
but he REF.3 interest-3SG also much PREP ART philosophy because
The abundant presence of subject pronouns in the constructed examples (2) and (3) is judged “totally unacceptable” by Fernández Soriano (1999:1227). However, in actual speech data one can easily come across examples where subject pronouns are used in a manner comparable to the examples (2) and (3), with three or more verbs with coreferential subjects. Thus consider example (4):

(4) *Bueno, pues yo quiero decir que...*  
Well, I want to say that.

Here the speaker expresses the first person singular subject pronoun with three consecutive verbs, without this affecting the “grammaticality” or “acceptability” of the utterance. Of course, the difference between example (4) and examples (2) and (3) is that the latter two exemplify narrative sequences with high discourse continuity. In such contexts, subject pronoun usage naturally differs from the use of first person singular verbs with stance-taking functions, as is the case in (4) where the speaker is introducing a narration of a personal experience. What causes the expression of subject pronouns in any particular case is typically a complicated matter and depends on many factors that may not be related to the acceptability or grammaticality of the utterance (cf. Fernández Soriano 1999) but rather depend on the local discourse context (see Travis & Torres Cacoullos 2012).

---

8 See Travis 2007 on the relation between genre and subject expression.
3.1.1. Contrast and emphasis

The received view in traditional grammar is that the use of subject pronouns signals either contrast or emphasis (cf. Butt and Benjamin 2004:130). Thus subject pronouns are usually considered obligatory in contexts where overt contrast is established between two referents, as in the constructed example (3):

(5) Yo hablo y tú escuchas.
I speak-1SG and you listen-2SG
‘I speak and you listen.’

Matos Amaral and Schwenter (2005) nevertheless observe that the expression of the subject is not strictly speaking obligatory if the clause contains another element such as an adverb referring metonymically to the referent of the subject. In addition, contexts such as the one exemplified in (5) are relatively rare in normal conversational data. Travis & Torres Cacoullos (to appear) examine the potential influence of contrast – as operationalized in three different ways in previous studies – in the expression of first person singular subjects in conversational Columbian Spanish. They find that even though subject pronoun presence may be attributed to contrast in some cases (namely in the case of “double contrast”, as defined in Myhill & Xing 1996), such cases are so rare in the data that the notion of contrast is not relevant for explaining the general patterns of subject pronoun usage.9

Standard definitions of contrast refer to a choice, either explicit or implicit, between two or more entities, as is the case in example (5) (see Umbach 2004 for a discussion). In contexts like the one exemplified by (4), it is not clear to what extent the concept of “contrast” is applicable, as no overt contrast is established between two possible subject referents, and there is thus no independent evidence for a contrastive reading apart from the presence of the subject pronoun. The frequent expression of pronominal subjects with mental verbs such as creer ‘think’ is sometimes attributed to the assumption that the use of such verbs is associated with an intention of contrast or “individualization” of the subject person (Fernández Soriano 1999:1236). However, following the definition of contrast given above, such uses of pronominal subjects are very seldom contrastive in the sense that a contrast would be established between two or

---

9 According to Myhill and Xing (1996: 314), double contrast implies a relation between two clauses that have two or more differing elements, as is the case in example (5): person X does A, person Y does B. In the case of single contrast (i.e. when there is only one different element in the clauses), subject pronouns need not be expressed. Even in the case of double contrast, subject pronoun expression is not obligatory.

As anecdotic evidence, consider the following excerpt from the dialogue of the film Amores Perros by Alejandro Inárritu:

Valeria, o abres o abro.
<name> or open-1SG or open-2SG
‘Valeria, either (I) open or (you) open.’

Repito que si no abres a las tres, voy a tirar esta puerta…
repeat-1SG that if NEG open-2SG at three go-1SG to break-INF this door
(I) repeat that if (you) don’t open [as I count] to three, (I) am going to break this door…’

This passage contains both a case of single contrast (o abres o abro) and double contrast (si no abres -- voy a tirar esta puerta) but no expressed subject pronouns.
more items (e.g. referents ‘I think so, not you’ or actions ‘I think so, I don’t suspect so’). Rather, the frequently occurring mental verb 
creer 
is used to signal the speaker’s epistemic stance (cf. Travis 2006; Posio 2011, 2012b).

As for the “emphasizing” function, it does not provide a satisfactory explanation of subject pronoun expression unless independent evidence of emphasis (in addition to subject pronoun presence) is provided. Thus in examples like (4), the expressed pronominal subjects can be considered emphatic in the sense that the speaker wants to draw hearers’ attention to herself and her personal experiences. Depending on how emphasis is defined, all subject pronoun usage that is not contrastive can be considered “emphatic” in the sense that it is not needed for the identification of the referent but rather serves some other functions. Thus the explanatory power of the concept of emphasis is weak and would require further clarification if the term is to be used. In Posio (2011, to appear) the notion of focusing of attention is suggested as an alternative way of conceptualizing the functions of subject expression in first person singular.

In conclusion, the terms “emphasis” and “contrast” have weak explanatory power unless properly defined. Given that it is impossible to emphasize or contrast a subject that is not overtly expressed in speech (Seo 2001: 70), the notions of emphasis and contrast may be applied to virtually all cases of subject pronoun usage. In consequence, there is an evident risk of circularity in the analysis if no independent evidence of the contrastive or emphatic reading is presented except the presence of the subject pronoun. In order to fully account for the relation of contrastivity and emphasis to subject pronoun expression, it would be crucial to study prosodic features such as intonation and stress. These factors have been largely ignored in previous research on subject pronoun expression. They also fall out of the scope of the present study, given that prosody is not encoded in the data examined (see Section 4.2).

---

10 Following Cornillie (2009), I use the term epistemic to refer to the speaker’s commitment to the contents of an utterance, distinguishing it from evidential which refers rather to the source of information.

11 The term “ellipsis” is sometimes also used to refer to the omission of subjects, especially in non-null subject languages. If “ellipsis” is understood to mean the omission of an element that is recoverable by the context (Crystal 2008:166), a prototypical case of subject ellipsis would be omitting the subject of the second member of a coordinated verb pair (e.g. he sings and Ø plays the piano). The relevance of such a notion in the description of Spanish and Portuguese is questionable, as the subjects of both the first and the second verb can be left unexpressed. In Spanish and Portuguese, subject ellipsis is not an independent phenomenon but rather pertains to the general tendency of leaving out subject pronouns that are coreferential with the previous subject. Coordinated verb pairs are very rare in the data analysed in the present work and in speech corpora in general. For instance, the Spanish data in the Corpus del Español contain only a handful of examples of coordinated first person singular verbs, typically in relatively fixed expressions, as the following:

\[
\text{desarrollo} \quad \text{un trabajo en el cual } \text{doy y recibo informació} \\
\text{develop-1SG a work where } \text{give-1SG and receive-1SG information}
\]

‘I develop a work where I give and receive information.’

CdE: CDEB033A.

Such cases are not considered separately from the rest of the data.
3.1.2. Grammatical persons and syncretisms

In various studies, the rate of subject pronoun usage has been shown to vary significantly between different grammatical persons. For instance, in the analysis Otheguy, Zentella & Livert (2007) grammatical person is chosen as the most significant variable conditioning the frequency of subject expression in New York Spanish. However, what causes the differences in the rate of subject expression in different persons is a complicated matter. Table 2 shows the frequency of subject pronoun expression in different grammatical persons in Enríquez’s (1984) and Cameron’s (1992) data from PS and Duarte’s (to appear) data from EP.¹²

Table 2. Frequency of expressed pronominal subjects in PS and EP

<table>
<thead>
<tr>
<th></th>
<th>PS (Madrid ;Enríquez, 1984: 348)*</th>
<th>PS (Madrid; Cameron 1992: 233)</th>
<th>EP (Duarte, to appear)***</th>
</tr>
</thead>
<tbody>
<tr>
<td>1SG</td>
<td>31.89% (3249/10185)</td>
<td>31% (N= 1010)</td>
<td>41% (173/421)</td>
</tr>
<tr>
<td>2SG</td>
<td>26.22% (134/511)</td>
<td>25% (N= 208)</td>
<td>39% (54/140)</td>
</tr>
<tr>
<td>3SG</td>
<td>13.34% (347/2601)</td>
<td>8% (N= 286)</td>
<td>25% (28/110)</td>
</tr>
<tr>
<td>1PL</td>
<td>10.40% (253/2431)</td>
<td>6% (N= 427)</td>
<td>32% (9/28)</td>
</tr>
<tr>
<td>2PL</td>
<td>11.11% (3/27)</td>
<td>-</td>
<td>28% (80/285)</td>
</tr>
<tr>
<td>3PL</td>
<td>14.12% (252/1794)</td>
<td>8% (N= 122)</td>
<td>26% (34/132)</td>
</tr>
<tr>
<td>usted**</td>
<td>72.70% (365/502)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ustedes</td>
<td>80.43% (37/46)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*For third person, masculine and feminine forms of the pronoun are counted together in this table. Generic uses of second person pronouns are excluded. **In PS, the pronoun usted (plural ustedes) is typically used as a formal address pronoun referring to the addressee in combination with third person verb forms. ***Duarte (to appear) does not specify whether the formal address pronouns você, o senhor etc. are counted as second or third person pronouns in the data; they refer to the addressee but are used with third person verb forms.

The data in Table 2 give a general picture of the tendencies related to subject pronoun expression in different grammatical persons in the corpora used by Enríquez (1984), Cameron (1992) and Duarte (to appear). Leaving aside the pronouns usted and ustedes in Enríquez’s (1984) data, the subject pronoun is expressed most often in first person singular in both PS and EP. Note that first person singular is also by far the most frequently used person form in both PS and EP data. The frequency of expression associated with first person singular has been attributed simply to the “egocentric nature of speech” (e.g. Claeys 2011:199) or the fact that “what we most like to talk about is ourselves” (Davidson 1996: 553). However, as discussed in Posio (2011, 2012a, 2012b, to appear), a great proportion of first person singular verbs used in speech have pragmatic and metacommunicative functions that set them apart from other grammatical persons (see Scheibmann 2001, Travis 2006). In other words, first person singular verb forms are not frequent just because speakers like to talk about themselves and the activities they are engaged in, but rather because many verbs occurring frequently in first person singular (e.g. ‘think’, ‘say’, ‘believe’, ‘know’) are used to express the

¹² The comparison of the percentages taken from different studies is problematic, as the categories are defined differently in each study and the corpora used are of different nature. For instance, Cameron (1992: 232) counts demonstrative pronouns (e.g. aquel, aquella) as third person subject pronouns and does not include the pronouns usted, ustedes or vosotros due to reduced number of occurrences.
speakers’ epistemic stance and to organize their discourse. Thus egocentrality should be understood as referring to the ways in which speakers construe their discourse, anchoring it to their subjective viewpoint (see Langacker 2008), not as a tendency to “talk about ourselves”.

Apart from the explanation according to which subject pronouns are used to mark contrast and emphasis, a traditional functional explanation argues that subject pronouns are used to disambiguate the subject person in the moods and tenses affected by syncretism between first and third person singular (viz. imperfect of indicative, conditional and subjunctive) or in dialects where the loss of postnuclear /s/ expands this syncretism to second person singular as well (see Table 3).

Table 3. Syncretic person marking in Spanish

<table>
<thead>
<tr>
<th>Imperfect of indicative</th>
<th>Conditional</th>
<th>Present of subjunctive</th>
<th>Imperfect of subjunctive</th>
</tr>
</thead>
<tbody>
<tr>
<td>cant-aba</td>
<td>cant-aria</td>
<td>cant-e</td>
<td>cant-ase</td>
</tr>
<tr>
<td>sing-IMP.1SG</td>
<td>sing-COND.1SG</td>
<td>sing-SUBJ.1SG</td>
<td>sing-SUBJ.IMP.1SG</td>
</tr>
<tr>
<td>cant-abas*</td>
<td>cant-arías*</td>
<td>cant-es*</td>
<td>cant-ases*</td>
</tr>
<tr>
<td>sing-IMP.2SG</td>
<td>sing-COND.2SG</td>
<td>sing-SUBJ.2SG</td>
<td>sing-SUBJ.IMP.2SG</td>
</tr>
<tr>
<td>cant-aba</td>
<td>cant-aria</td>
<td>cant-e</td>
<td>cant-ase</td>
</tr>
<tr>
<td>sing-IMP.3SG</td>
<td>sing-COND.3SG</td>
<td>sing-SUBJ.3SG</td>
<td>sing-SUBJ.IMP.3SG</td>
</tr>
</tbody>
</table>

*Syncretic in dialects with loss of postnuclear /s/.

Although the hypothesis that syncretic person marking is compensated by subject pronoun use found support in Hochberg (1986), it has later been refuted in several studies such as Ranson (1991) and Cameron (1992, 1993). The idea that null subjects occur only in languages with non-syncretic person marking (known as Taraldsen’s generalization: see Gilligan 1987) is widespread but lacks cross-linguistic validity. On the one hand, there are languages like German and Icelandic that have distinctive subject marking on verbs but nevertheless require subject pronouns to be used. On the other hand, subjects and other arguments may be left unexpressed also in languages like Chinese and Japanese that lack inflectional person marking altogether. Avoiding the eventual ambiguity caused by the person syncretism is unlikely to motivate subject pronoun usage especially in spoken language, where the referent of the subject is usually made clear by the context.\(^{13}\) The influence of personal syncretisms has nevertheless been considered crucial in accounting for the differences between European and Brazilian Portuguese (see Section 3.2.1).

Although subject pronouns may naturally be used also for disambiguating purposes, in practice disambiguating seems to be a minor function which is often difficult to distinguish from other uses (Posio 2012a). Note also that subject pronoun usage is by no means limited to the syncretic forms. Although several studies suggest that the tenses and moods with syncretic person marking are associated with frequent expression of subject pronouns, this phenomenon may also be related to the pragmatic uses of such moods and tenses in discourse (see Silva-Corvalán 1997, 2001 and Section 3.1.6).

\(^{13}\) However, see Paredes Silva (1993) on ambiguity effects in written Brazilian Portuguese.
3.1.3. Sociolinguistic and areal variation

The expression of pronominal subjects constitutes a frequent research topic in sociolinguistic studies (see Flores-Ferrán 2007 for an overview). Subject expression is often regarded as a discrete binary phenomenon that does not affect the truth value of utterances and that is conditioned by factors that may be operationalized for a sociolinguistic analysis (Erkeli Guy, to appear) and that are assumed to be independent of each other (Otheguy, Zentella & Livert 2007:773). Several studies have revealed dialectal differences in the frequency of subject pronoun expression: the highest rate of expressed subject pronouns is found in Caribbean Spanish, especially in Puerto Rico (e.g. Hochberg 1986, Morales 1989, Cameron 1993, Flores-Ferrán 2004, Otheguy, Zentella & Livert 2007; but cf. Claes 2011).

The frequent expression of subject pronouns in Caribbean Spanish has been connected to a lower rate of postposed subjects in that variety (Morales 1989; cf. also Posio 2012a). Contrastive studies on Spanish varieties have focused mostly on Latin American Spanish: less effort has been put to the analysis of PS and there is thus little information on eventual dialectal differences within the Peninsular variety (but cf. Ranson 1991 on Andalucian Spanish).

Although some studies have found evidence to support a connection between the age or educational background of the speaker and subject pronoun expression (e.g. Ávila-Jiménez 1996, Flores-Ferrán 2002), typical sociolinguistic factors such as gender, age, education, socioeconomic status and class of the speaker have in general showed relatively little influence on subject pronoun usage (see Enríquez 1984; Bentivoglio 1987:57-58; Cameron 1992; Silva-Corvalán 1997:122, 2001:133, Otheguy, Zentella & Livert 2007:778). This may be due to the fact that sociolinguistic studies typically have not focused on local contexts of subject expression (such as specific verbs or constructions) but rather subject expression in general. For instance, Aijón Oliva & Serrano (2010) find that speakers’ gender and profession co-vary along with the expression of pronominal subjects in the construction (yo) creo ‘I think’, although such differences are not manifested in the general usage of subject pronouns apart from that specific construction. Socially or areally conditioned variation is left out of the scope of the present work, but in future research studying the connection of specific discourse patterns and socio-situational factors may provide more interesting results.

In variationist sociolinguistic research design, great care is typically put into defining the envelope of variation within which variation may occur. In the case of the expression of pronominal subjects, contexts falling outside the envelope of variation include e.g. zero-argument verbs (e.g. weather verbs and the existential predicate hay ‘there is’), subject-headed relative clauses, clauses with NP subjects (Otheguy, Zentella & Livert 2007:775-778, Adli 2011:229). Since one of the purposes of the present work is to investigate subject pronoun usage in formulaic constructions, i.e. in contexts where it is highly conventionalized, it would be counter-productive to exclude all seemingly conventionalized sequences consisting of verbs and subject pronouns or verbs without

---

There has been debate over the possible influence of English in contact varieties of Spanish, but so far the evidence in favour of such an effect has been contradictory (see Flores-Ferrán 2007: 643-644 and references therein and Travis & Torres Cacoullos, to appear). Otheguy, Zentella & Livert (2007) find that Spanish-speaking South American immigrants in New York tend to increase their subject expression rate as they spend more time in contact with English, but this might also be due to more intense contact with speakers of Caribbean varieties of Spanish in New York.
subject pronouns from the scope of the analysis. In addition, in many cases the “obligatory” and “variable” contexts of subject pronoun expression or placement do not fall into two discrete categories. For instance, in Posio (2012a) focusing on postverbal pronominal subjects, one might consider excluding interrogative clauses with preverbal question words from the envelope of variation because in such clauses subjects are required to be postverbal. However, given that the conditions of postverbal placement are assumed to be gradual, interrogative clauses are not to be seen as fundamentally different from other contexts where subjects are placed postverbally. The only difference is that postverbal placement of the subject has grammaticalized with question words and thus has become obligatory, whereas in other cases it is only a tendency (see Posio 2012a).\footnote{In addition, one can easily observe by looking at the data that postverbal placement of the subject after question words is not an invariable rule in Spanish or Portuguese and exceptions do occur.}

3.1.4. Information structure and switch reference

The term information structure refers to the relationship between “the speaker’s assumptions about the hearer’s state of knowledge and consciousness at the time of an utterance” and the way in which the speaker formulates his or her utterances guided by these assumptions (Lambrecht 1994: xiii). Information structure is argued to account for the speakers’ choice between referring expressions and is thus expected to play a role in the variation in subject pronoun expression and placement (see e.g. Bentivoglio 1983, 1993; Morris 1998). Theories dealing with anaphora and the choice of referring expressions consider personal pronouns (as opposed to noun phrases) as markers of a high level of accessibility (Ariel 1994, 2001) or givenness (Gundel, Hedberg and Zacharski 1993; Taboada 2005) of the referent. In languages with variable subject expression, the use of subject pronouns instead of null anaphora marks lower accessibility of the referent (Ariel 1994:30, 2001:31). In order for a null anaphor to be used, the referent is required to be “at the current center of attention” (Gundel, Hedberg and Zacharski, 1993:279).

Theoretical approaches to accessibility or givenness (as well as empirical research on the topic) tend to focus on anaphoric third person referents. It is unclear to what extent the theory accounts for the variation between subject expression and omission in first and second person. In the case of local persons (viz. first and second person singular), the referents are present in the communicative situation\footnote{First and second person singular pronouns typically refer to the speaker and the addressee, although especially the second person singular pronoun can be used for generic reference (see Fernández Soriano and Táboas Baylín, 1999:1723).} and thus need not be activated in the short-term memory of the speaker and addressee. According to Ariel (2001:32), first and second person inherently display a higher level of accessibility than third person referents. Therefore it is likely that information structure does not explain the usage of first person pronouns, at least to the same extent as argued for third person referents. It is argued in Posio (2011, 2012a, to appear) that in the case of first person singular, the expression of the subject pronoun signals greater focusing of attention on...
the referent of the subject, whereas not expressing the pronoun may be associated with a higher level of attention focused on other elements.

In empirical research on Spanish subject expression, information structure has been partially operationalized through the notion of switch reference. The use of subject pronouns has been shown to correlate with switch reference in a number of studies (e.g. Silva-Corvalán 1982; Enríquez 1984; Cameron 1992, 1993, 1995, 1997; Hurtado 2005). Simply put, the notion of switch reference captures the fact that a pronoun is more frequently used instead of a null subject if the subject is not coreferential with the previous subject. Switch reference is an important factor explaining the usage of subject pronouns. It is related to information structure in the sense that new-mention, lowly accessible subjects are typically switch reference subjects whereas given or highly accessible subjects are same reference subjects. As an explanatory factor, switch reference is thus fully compatible with the principles of information structural accounts, such as Ariel’s (1994, 2001) Accessibility Theory or Centering theory (Gundel, Hedberg & Zacharski 1993). However, if switch reference was the only relevant factor, there would presumably be no major differences between different verbs or semantic roles in terms of subject pronoun usage. It is also an interesting question to what extent the effects of switch reference are similar in the anaphoric third person and in the deictic persons.

3.1.5. Semantics and pragmatics of subject expression

Since semantics and pragmatics of subject expression are in the main focus of this dissertation, relevant research literature is discussed in all articles. Thus instead of giving an extensive review of the previous literature here, this section merely summarizes the approach adopted in previous studies and their main findings. For a more detailed, critical overview of the previous research, see Posio (2011:779-782).

It has been observed in several studies that different verb lexemes or semantically defined verb classes have different subject pronoun expression rates (e.g. Enríquez 1984; Bentivoglio 1987; Morales 1997; Miyajima 2000; Hurtado 2005). In most previous studies on the connection between verb semantics and subject expression, verbs have been divided into semantically motivated categories and the research has focused on comparing the frequency of subject pronoun expression between these categories. The results of different studies are not directly commensurable, as the criteria for data selection and the categories used vary from study to study. In some studies different grammatical persons are analyzed separately while in other studies data from all or some grammatical persons are grouped together. As noted by Cameron (1992: 45–49), this is problematic because verb semantics seems to affect systematically only first person singular, or at least it is not clear whether the influence is similar in all grammatical persons.

Table 4 presents the categories used in Enríquez (1984) and the frequency of subject pronoun presence in each category.

---

17 See Adli (2011:220) on the differences between information structural accounts and switch reference accounts of subject pronoun expression.
Table 4. Verb categories in Enríquez (1984)

<table>
<thead>
<tr>
<th>Category</th>
<th>Subject Pronoun Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Verbs that express an opinion or judgement of the speaker (like creer ‘believe, think’, considerar ‘consider’, imaginar ‘imagine’):</td>
<td>54.45 %*</td>
</tr>
<tr>
<td>2. Verbs that express mental activity (like saber ‘know’, querer ‘want’, aprender ‘learn’, imaginar ‘imagine’):</td>
<td>28.17 %</td>
</tr>
<tr>
<td>3. Stative verbs that do not express a dynamic process exercised by the subject (like ser ‘be’, estar ‘be’, tener ‘have’, saber ‘know’, vivir ‘live’):</td>
<td>22.94 %</td>
</tr>
</tbody>
</table>

*The percentages are calculated on the basis of data from all grammatical persons.

Enríquez (1984:244-245) attributes the higher frequency in category 1 to a tendency of contrasting\(^{18}\) the subject referent of verbs that are used to express personal opinions with other referents. Such verb classifications are used e.g. by Bentivoglio (1987), Morales (1997) and Hurtado (2005), with similar results. In all studies, the category including “mental” or “cognitive” verbs scores the highest frequencies. However, this category may either include or exclude other categories such as modal verbs (Bentivoglio 1987) or communicative verbs (Morales 1997), depending on the study. Thus the only common finding seems to be that the loosely defined category of verbs expressing mental activities seems to favor subject pronoun usage. However, semantically related verbs included in a category of “mental” or “psychological” verbs have been shown to have highly divergent distribution across persons and tenses as well as different usage patterns (Travis 2006). Further problems related to previous research based on semantic verb classes are discussed in Posio (2011, 2012a, to appear).

Several pragmatic factors have also been suggested to account for the usage of subject pronouns with certain verb lexemes. Davidson (1996) argues that subject pronouns are used to trigger speech act readings of some verbs or disambiguate epistemic parentheticals is related to such frequently occurring verbs as creer ‘think, believe’ or saber ‘know’. While for example the forms creo ‘I think’ and no sé ‘I don’t know’ are frequently used in speech as epistemic parentheticals with no truth-functional meaning (i.e. they do not refer to thinking or knowing anything in concrete but function more as epistemic parentheticals), the same verb forms used with a subject pronoun are argued to have a more concrete meaning (Davidson, 1996:559). In general, the expression of subject pronouns is argued to add “pragmatic weight” to the utterance (Davidson, 1996:551; cf. Posio 2011:781–782, 786, 793–794 for a discussion of the notion). As for first person singular pronouns, Stewart (2000) also argues that their use with verbs of cognition can be considered as a way of marking the speaker’s opinions as their own, not necessarily shared by the addressee. Thus expressions such as yo creo que ‘I think that’ can be thought of as “hedges” in the sense of Brown & Levinson (1987).

\(^{18}\) Enríquez (1984:245) uses the expression “el deseo de contraposición de personas”, i.e. ‘the desire of contrasting/comparing persons’. 
3.1.6. Semantic and pragmatic salience or focusing of attention

While information structural accounts usually define salience as the accessibility of the referent due to previous mentions in text, in semantic approaches the notion of salience or prominence\(^{19}\) is used to refer to the foregrounding of an element with respect to other elements in an event or situation (Langacker 2008:66). Typically, subjects are said to be salient with regard to direct objects or oblique participants (Givón 1994) as in the prototypical case they are volitional instigators of actions (Dowty 1991). Salience, in the latter sense, may also be manipulated for pragmatic purposes of the speakers: elements of an utterance may be foregrounded or backgrounded with respect to other elements by linguistic means. For instance, in Posio (2011, to appear) it is argued that the expression of pronominal subjects in first person singular serves to foreground the subject referent while omitting the pronoun functions as a grounding device. In Posio (2012a) it is argued that postverbal placement of the subject also serves a grounding function, especially with those subject pronouns that are frequently expressed even when the referent is not foregrounded (such as the Spanish address pronoun usted and Portuguese a gente; cf. Section 3.2 and Posio 2012b).

Several studies have pointed at more frequent subject expression in the imperfect of indicative, conditional and subjunctive than in other tenses and moods. While this has been regarded as a case of ambiguity resolution (it is precisely in those tenses and moods where first and third person singular have syncretic forms), Silva-Corvalán (1997:127, 2001) suggests that the more frequent expression of pronominal subjects is related to the functions of different tenses and moods in discourse and the focusing of attention on either the subject or the action expressed by the verb. She argues that subject pronouns are expressed most frequently in those tenses and moods that are related to backgrounded, irreal, non-dynamic and non-assertive narrative functions, and the expression of the subject serves to draw attention away from the action or the situation. In the perfective past tenses the action depicted is more prominent, foregrounded, real and assertive, and the omission of subject pronouns allows the action to be put on the fore.\(^{20}\) In Posio (2011, to appear) it is argued that along with backgrounding and foregrounding as narrative functions, high level of clausal transitivity (in the sense of Hopper & Thompson 1980) may also reduce the rate of subject pronoun presence, as more attention is focused on the affected Patient or on the dynamic action than in clauses characterized by lower transitivity.


---

\(^{19}\) The terms “prominence” and “salience” are often used interchangeably (cf. Langacker 2008:66) or one of the terms is used in the definition of the other (cf. Chiarcos, Claus & Grabski 2011:2).

\(^{20}\) The terms backgrounding and foregrounding are usually associated with narrative functions (see Wårvik 2004), but in the present study they rather refer to the linguistic grounding of clausal participants and verbal action in line with the cognitive Figure vs. Ground distinction (Wallace 1982). Typical properties of backgrounded verbal categories are stative, durative, nonperfective, incompleted, non-present or remote, irreal, negative, intransitive and accidental (Wårvik 2004:108). These properties of the predicate correlate with more frequent subject expression, given that more attention is focused on the subject referent as opposed to the action or state expressed by the verb (Posio 2011). On the contrary, when the action itself is salient – i.e. eventive, punctual, perfective, completed, real, affirmative, transitive and deliberative (Wårvik 2004:108) – and involves an affected Patient, the expression of the pronominal subject tends to be less frequent (Posio 2011).
with a special emphasis on the creation of different communicative styles by using different strategies of subject expression. Their starting point is the hypothesis that subject expression (in either preverbal or postverbal position) and subject omission reflect two different conceptualizations of the relation between the participants of the depicted event (Aijón Oliva & Serrano 2010:10). The salience of the participants is defined with reference to two components: prominence or activation of the referent in the discourse context and focalization or informativity of the referent. Subjects associated with a high degree of activation are omitted, whereas expressed subjects are more informative than omitted ones (Aijón Oliva & Serrano 2010:10). The model is summarized in Table 5.

<table>
<thead>
<tr>
<th>Type of subject</th>
<th>Example</th>
<th>Prominence or activation</th>
<th>Focalization or informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>omitted</td>
<td>vengo</td>
<td>+ prominent/activated</td>
<td>- focal/informative</td>
</tr>
<tr>
<td>preverbal</td>
<td>yo vengo</td>
<td>+/- prominent/activated</td>
<td>+/- focal/informative</td>
</tr>
<tr>
<td>postverbal</td>
<td>vengo yo</td>
<td>- prominent/activated</td>
<td>+ focal/informative</td>
</tr>
</tbody>
</table>

Postverbal subjects are considered to be more informative – due to the rarity or unexpectedness of postverbal placement – but less prominent than subjects expressed in the preverbal position (Serrano 2012: 109-131). The unexpressed subject represents the highest level of prominence. In a study focusing on the expression of the 1SG subject yo with the verb form creo, Aijón Oliva and Serrano (2010: 35) conclude that in this particular case the presence of the subject pronoun is related to a more subjective relation of the speaker to the propositional contents of the utterance, whereas the absence of the pronoun is connected to a more objective stance. The two strategies of subject marking thus participate in the processes of subjectivization and objectivization of the utterance, respectively, and work in the creation of a more subjective or objective communicative style.

The two notions of salience discussed above – information-structural salience and semantic salience or prominence – should not be confused. In the current work, the term salience is used in the former sense, whereas the latter is conceptualized through the notion of focus of attention (Posio 2011, to appear). This being said, the current work focuses more on the semantic salience of participants in events and the salience of participants carrying different semantic roles. Since the focus is on non-anaphoric subjects, measuring the topicality or accessibility of the referent is not among the main concerns of the work. First person referents – especially first person singular – differ

---

21 In Posio (2012a) it is argued that the postverbal placement is not always related to informativity as hypothesized in the model proposed by Aijón Oliva & Serrano (cf. Table 5) but may also serve a backgrounder function.

22 Note that “prominence” refers here to a prominent mental representation of the referent of the subject pronoun, caused e.g. by previous mentions in the discourse. The distinction between the terms “subject” and “referent of the subject” are crucial, as the prominence of the referring expression (such as a subject pronoun) is typically inversely proportional to the prominence of the referent (cf. Ariel 2001). In other words, higher prominence of the referent is reflected in lower prominence of the linguistic encoding given to the referent and vice versa.
from anaphoric referents in that they are always accessible due to their presence in the communicative situation. In the anaphoric third person, using a null subject requires that the referent is salient and recognizable to the addressee. Given that first person singular referents are inherently accessible, null subjects may be used more freely than in third person. While null subjects are associated with high saliency in anaphoric persons, the use of subject pronouns in first person singular is argued to signal increased focusing of attention on the subject referent (see Posio 2011, to appear).

### 3.1.7. Priming effects

In addition to switch reference and topicality, the psycholinguistic effect called perseveration (Cameron & Flores-Ferrán 2004) or priming (Travis 2007) has been connected with subject pronoun expression and omission in previous research (see Cameron 1994, 1995; Flores-Ferrán 2002; Cameron & Flores-Ferrán 2005; Travis 2007). The notion of priming originates in experimental psychology: in linguistics it is used to refer to inconscious repetitions of such sequences that the speaker has been recently exposed to (Gries 2005). Such repetition is supposed to be spontaneous and functionally unmotivated (Cameron & Flores-Ferrán 2004:42). Travis (2007), who focuses on first person singular subjects, defines subject pronoun priming as a phenomenon “whereby a pronominal mention favors a subsequent pronominal mention, and an unexpressed mention favors a subsequent unexpressed mention to a statistically significant degree” (Travis 2007:131). This effect is assumed to be independent of accessibility and switch reference and may actually work in the opposite direction: while the model of switch reference predicts that subjects functioning as continuous topics in discourse are referred to by null subjects, priming might rather provoke the expression of pronominal subjects with subsequent mentions of the same subject referent.

Although the current work does not focus on priming effects on subject pronoun usage, it is worth noticing that priming may affect not only the expression or omission of subject pronouns but also the use of specific verb forms or lexemes, or formulaic sequences including verbs with or without subject pronouns. Such formulaic sequences as *yo creo* ‘I think’ or *no sé* ‘I don’t know’ are argued to be stored and retrieved holistically: thus it is to be expected that the whole sequence is affected by priming. Another potentially controversial issue related to priming is that while it is defined (or assumed to be) spontaneous and functionally unmotivated, in practice it may be difficult to distinguish from structural or lexical repetition that serves a specific function. Thus consider example (6) from EP:

---

23 However, first person singular verb forms or pronouns do not always refer to the speaker, e.g. when occurring in constructed dialogue or reported speech.
And the man himself, won’t he feel that he’s losing his virility?”

‘No. That, I think it doesn’t have to be like that.

I think not. Then, why would he lose it?

On the contrary, if he, if he’s going to use things that he can feel better, much..., in conclusion,

I think not. What is it, why would he lose his virility?

I think it doesn’t make sense.’

The repetition of the sequence acho que não (with modifications at the two last lines of the turn) might be caused by unconscious and pragmatically unmotivated repetition of a previously heard sequence (i.e. a priming effect). However, it is also possible that in (6) the speaker uses repetition as a stylistic device with pragmatic functions, i.e. to manifest her disagreement with the previous speaker. Of course, such repetition may also be unconscious even if the speaker uses it for a communicative purpose.

3.1.8. Frequency effects

One of the main findings of this dissertation is that subject pronoun usage is affected by the relative frequency of verb forms and verb lexemes (see Posio to appear). The eventual influence of frequency (see Bybee & Thompson 1997) on subject pronoun expression has not been taken into account in previously published research on subject pronoun usage in Spanish or Portuguese except for the recent studies by Erkel & Guy (to appear) and Silveira (2011; see Section 3.2.1). The paper by Erker and Guy (to appear) examines precisely the influence of frequency on the presence or absence of subject personal pronouns in New York Spanish. Their results, discussed in this

---

24 The three frequency effects discussed in Bybee & Thompson (1997) are the reduction effect (frequently used items tend to suffer phonological reduction), the conserving effect (high frequency of use cases irregularities to resist change) and the productivity related to high type frequency (items that have a high type frequency, e.g. because they may be used in a large amount of different contexts, are likely to be more productive than items with a low type frequency).
subsection, are consistent with my findings from Peninsular Spanish and European Portuguese.

One of the hypotheses tested in Erkel and Guy (to appear) is that if there were an ongoing change in Spanish from less frequent to more frequent subject pronoun expression (as evidenced by Brazilian Portuguese and French; see Section 3.2.1), null subjects should be archaic and, in accordance to the conserving effect of frequency (Bybee & Thompson 1997), they should be more frequent with the most frequently occurring verbs forms. However, this is not the case in Erkel and Guy’s (to appear) data, and there is no other evidence of an ongoing change leading into more frequent subject pronoun expression in Spanish (unlike in Brazilian Portuguese; see Section 3.2.1). A second hypothesis is that frequent verb forms commonly occurring without subject pronouns come entrenched with the null subject collocation, while forms commonly occurring with expressed subject pronouns come entrenched in that association. The actual findings of Erkel and Guy (to appear) are in line with the second hypothesis.

Examining a corpus of 5,000 clauses from 12 speakers of Spanish residing in the United States, Erkel and Guy (to appear) find that the frequency of the verb form is not a significant factor by itself: among the frequently occurring verb forms, there are forms highly favoring the expression of the subject pronoun (e.g. *sabes* ‘you know’, with 92% of subjects expressed) and forms highly disfavoring the use of subject pronouns (e.g. *digo*, ‘I say’, with 12% of the subjects expressed). A closer examination of the different factors included in the analysis reveals that frequency interacts with the other explanatory factors. First, verb forms in the corpus are classified as frequent (in which case a single form constitutes at least 1% of the data) and infrequent (all other verbs). The group of frequent verb forms consists of 13 verb forms that make up 22.8% of the corpus. Comparing the groups of frequent and infrequent verbs, Erkel and Guy (to appear) demonstrate that the factor groups verb type (mental, stative or external activity verb), morphology (irregular vs. regular verb) and grammatical person have a significant effect on the frequency of subject pronoun usage only among the frequent verbs forms. The factor groups tense/mood/aspect (imperfect, present, preterite) and switch/same reference show an effect in both groups, but again the effect is stronger in the group of frequently occurring verbs. Erkel and Guy (to appear) relate their findings with the entrenching effect of high frequency:

High frequency verb forms are the ones that speakers have more experience with and hear more often from others. […] We propose, therefore, that frequency affects linguistic variation not directly, but indirectly, as a limitation on the sufficiency of evidence. […] Below some frequency threshold, items are too rare to formulate rich representations that include collocational information. Above the threshold, language users have sufficient information about each lexical item to individuate them with respect to collocations and syntactic operations.

The study of Erkel and Guy (to appear) thus deals partially with the same questions as the present study (see especially Posio to appear), although the methodology chosen is different. Erkel and Guy’s study is purely quantitative research based on multivariate analysis, while my research is more qualitatively oriented. Thus the scope of generalizations made and the explanations offered are of different nature. While Erkel and Guy refer to the entrenching effect of frequency, they do not present an account of
what exactly causes certain verb forms to be used frequently. Looking at the 13 verb forms in the category of frequently occurring verbs in Erkel and Guy’s data, most of the forms are first person singular forms and most of them are being focused on in the articles of the present study. The frequent verb forms in Erkel & Guy’s data are the following: creo ‘I think’, sé ‘I know’, digo ‘I say’, tengo ‘I have’, sabes ‘you know’, ves ‘you see’, estaba ‘you/(s)he was’, estoy ‘I am’, tenía ‘I/(s)he had’, era ‘I/(s)he was’, soy ‘I am’, fui ‘I was/went’, and es ‘(s)he is. For instance, Posio (2011) analyzes precisely the Spanish forms creo, sé, digo, tengo, sabes, ves, estoy, and soy, and Posio (to appear) also discusses the most frequently occurring first person singular forms creo, digo, sé and their equivalents in the Portuguese data. It is argued in Posio (2011, to appear) that the fact that these verb forms occur frequently permits the fixation of subject pronoun presence/absence and position with these forms and thus the creation of formulaic sequences. Thus high frequency not only permits rich memory representations of the factors affecting subject pronoun usage in general, but also allows for verb-specific patterns of subject pronoun usage to emerge (Posio 2011, 2012a, 2012b, to appear).

3.2. Research on Portuguese

Compared to Spanish, subject pronoun usage in Portuguese has received little attention. Most of the research has focused on Brazilian Portuguese, which is argued to be gradually developing from a null subject language into a language where pronominal subjects are normally if not obligatorily expressed (see Lira 1982, 1996; Duarte 1993, 1995, 1998, 2000; Barbosa, Duarte & Kato 2001, 2005; Cyrino, Duarte & Kato 2000; Modesto 2000; Silveira 2011). In studies focusing on Brazilian Portuguese, the European variety is presented as a prototypical null subject language, comparable to Spanish or Italian. Given the typological proximity of Peninsular Spanish and European Portuguese (see Section 1.3.), one might indeed expect to find similar patterns of subject pronoun expression in the two languages.

Since the semantic, pragmatic or cognitive aspects of subject pronoun usage in Portuguese have not received as much attention in previous research as they have in Spanish, the previous research is not presented under the headings used for Spanish in Section 3.1. Rather, Section 3.2.1 focuses on the research carried out on Brazilian Portuguese (henceforth BP). Studies comparing BP and EP are reviewed in section 3.2.2, which also discusses some differences that are known to exist between Portuguese and Spanish regarding subject pronoun usage.25

25 Spanish and Portuguese subject pronoun expression has been previously compared in Soares da Silva (2006) who contrasts his own frequency data from Buenos Aires and Madrid with Duarte’s (1995) data from Brazilian and European Portuguese and data from Mozambican Portuguese. The result of the comparison is that the high frequency of expressed subject pronouns differentiates Brazilian Portuguese from the other varieties. Unfortunately the data used are not directly comparable and no raw frequencies are given to permit the evaluation of the validity of the results. Soares da Silva (2006) does not study the the use of subject pronouns with different verb forms or lexemes.
3.2.1. Expression of pronominal subjects in Brazilian and European Portuguese

Several studies focusing on the expression of pronominal subjects in BP (e.g. Duarte 1993, 1995, 1998, 2000; Barbosa, Duarte & Kato 2001, 2005; Cyrino, Duarte & Kato 2000; Modesto 2000) claim that BP is gradually losing its null subjects and that expressing the subject pronoun is the normal or “unmarked” option in most contexts. BP and EP have been explicitly contrasted in Duarte (1995) and Barbosa, Duarte & Kato (2001, 2005). These studies have been carried out in the generative framework and the theoretical discussion is, to a great extent, incommensurable with the functional perspective adopted in the present study. A recent exception is Silveira’s (2011) dissertation which focuses on the expression of pronominal subjects in first, second and third person singular taking into account semantic and lexical factors and frequency effects. In Silveira’s (2011) data from spoken BP, the frequency of expressed subject pronouns is 66% in first person singular, 54% in second person singular and 48% in third person singular (with human referents).

In a diachronic study on subject pronoun expression in BP, Duarte (2000) shows that subject expression gradually becomes more frequent overt time. The data are theatrical dialogues from Brazilian Portuguese plays from different periods. The increase in the presence of expressed pronominal subjects correlates with a change in the subject marking paradigm, divided into three phases in Duarte’s (2000) model, as shown in Table 6. Paradigm 1 is “functionally rich” (Roberts 1993), given that it contains only one null morpheme and only one syncretism, i.e. that between second and third person singular. At this stage, expressed subjects “seem to be used for emphasis or focus, as is usual in pro-drop systems” (Duarte 2000: 20). Paradigms 2 and 3 show a weakening of the functional richness assumed to be needed for subject omission to be possible. As to be expected, the frequency of subject expression rises gradually with the change from one paradigm to another. The change is most notable in first and second person, while in third person null subjects resist longer: this is attributed to the fact that in third person the referent can be identified by the presence of a NP subject in previous context (Duarte 1995, 2000:20).

Using theatrical dialogue to estimate the timing of diachronic changes can naturally be criticized for many reasons. In addition to the linguistic change itself, many extralinguistic factors may intervene in the change. What is more, not all playwrights have the ambition of faithfully representing the way people speak at the time of writing the play, and even if a statistically relevant change has taken place in actual speech, it usually takes a considerable time before it is reflected in writing. However, theatrical dialogues are one of the few available sources for the study diachronic changes in

---

26 In discussing the previous research, I will adhere to the principle that empirical results and statements made about language use are not theory-specific and can therefore be discussed or challenged within another theoretical framework. In contrast, I will not discuss the assumptions and claims formulated about the syntactic descriptions or the generative theory itself.

27 Although Duarte (2000) does not include the first person plural form (i.e. the subject pronoun nós and the verb forms ending in -mos) in her Paradigm 3, first person plural has not been completely lost from modern Brazilian Portuguese. According to Travis & Silveira (2007), it remains in use especially with the most frequently occurring verbs in forms such as temos ‘we have’ and somos ‘we are’ due to the conserving effect of frequency (see also Posio 2012b).
spoken language, and the evidence of the change seems rather robust even if the actual timing of the change may be difficult to estimate.

Table 6. Person marking paradigms in Brazilian Portuguese plays (Duarte 1993, 2000:19)*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1SG: eu</td>
<td>am o</td>
<td>am o</td>
<td>am o</td>
</tr>
<tr>
<td>2SG: tu</td>
<td>am a s</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2SG: você</td>
<td>am a</td>
<td>am a</td>
<td>am a</td>
</tr>
<tr>
<td>2SG: o senhor, a senhora</td>
<td>am a</td>
<td>am a</td>
<td>am a</td>
</tr>
<tr>
<td>3SG: ele, ela</td>
<td>am a</td>
<td>am a</td>
<td>am a</td>
</tr>
<tr>
<td>1PL: nós</td>
<td>am a mos</td>
<td>am a mos</td>
<td>-</td>
</tr>
<tr>
<td>1PL: a gente</td>
<td>-</td>
<td>am a</td>
<td>am a</td>
</tr>
<tr>
<td>2SG: vós</td>
<td>am a is</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2SG: vocês</td>
<td>am a m</td>
<td>am a m</td>
<td>am a m</td>
</tr>
<tr>
<td>2PL: os senhores, as senhoras</td>
<td>am a m</td>
<td>am a m</td>
<td>am a m</td>
</tr>
<tr>
<td>3PL: eles, elas</td>
<td>am a m</td>
<td>am a m</td>
<td>am a m</td>
</tr>
</tbody>
</table>

Frequency of pronominal subjects in theatrical dialogue


*The percentages given here include the formal second person address forms o senhor, a senhora, os senhores, as senhoras that are excluded from Duarte (2000)

An interesting question concerning Duarte’s (1995, 2000:20) results is why the change is so drastic precisely in first person singular where the expression rate rises from 31% in 1845 to 82% in 1992. First person singular is the only person not affected by syncretisms in the present tense in any of Duarte’s three paradigms. Third person singular, on the contrary, is the person most affected by syncretisms or impoverished person marking, yet it shows the highest proportion of null subjects (Silveira 2011: 34, 46). Even if compensating for the ambiguity caused by syncretisms is considered to motivate the increase in subject pronoun frequency over time in BP, it is unclear why the change is most abrupt in first person singular which is not affected by the new syncretisms as opposed to third person which is directly affected by the syncretisms (cf. Duarte 1995, 2000).

In the case of first person singular, it is not plausible that the identification of the referent be more difficult than in third person due to the lack of expressed antecedents (cf. Duarte 2000:20), as first person singular referents are always present in the communicative situation and therefore inherently more accessible than anaphoric referents (Ariel 2001:32). Note also that functional compensation of syncretic (i.e. ambiguous) person marking as an explanation of subject pronoun expression has been largely refuted in studies focusing on Spanish (see Section 2.1.2.).

Compared to BP, person marking in EP resembles Duarte’s (2000) paradigms 1 and 2. EP conserves second person singular but has lost second person plural (apart

---

28 Recall that Portuguese, similarly to Spanish, has syncretic person marking in first and third person singular in the imperfect of indicative, conditional and subjunctive. These syncretisms remain unchanged during the three periods distinguished by Duarte (2000).
from some dialects). In Table 7, the EP paradigm is contrasted with the Peninsular Spanish paradigm. The new first person plural pronoun *a gente* is used along with the first person plural form ending in -*mos*: both seem to be equally frequent in speech although there is lots of variation among individual speakers (Pereira 2003, Posio 2012b). While in BP *a gente* has replaced the older first person plural form in most contexts (Travis & Silveira 2009), in EP it conserves its older impersonal use (Posio 2012b). All in all, the most salient difference between BP and EP person marking paradigms is that EP has one distinctive person form more than BP: the second person singular. The difference between EP and PS, then, boils down to the fact that the 2PL form remains in use in PS but not in EP. What is crucial to the present study, then, is that in first person singular the number of syncretisms is the same in EP and PS. Thus the differences found between PS and EP cannot be plausibly attributed to a higher number of syncretisms in one of the languages.

Table 7. Comparison of EP and PS person marking paradigm

<table>
<thead>
<tr>
<th>European Portuguese</th>
<th>Peninsular Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>1SG: <em>eu</em></td>
<td>1SG: <em>yo</em></td>
</tr>
<tr>
<td>2SG: <em>tu</em></td>
<td>2SG: <em>tú</em></td>
</tr>
<tr>
<td>2SG: <em>você</em></td>
<td>2SG: <em>usted</em></td>
</tr>
<tr>
<td>2SG: <em>o senhor,</em></td>
<td>2SG: <em>am a</em></td>
</tr>
<tr>
<td>3SG: <em>ele,</em></td>
<td>3SG: <em>él,</em></td>
</tr>
<tr>
<td>1PL: <em>nós</em></td>
<td>1PL: <em>nosotros,-as</em></td>
</tr>
<tr>
<td>1PL: <em>a gente</em></td>
<td>1PL: <em>am a</em></td>
</tr>
<tr>
<td>2PL: <em>vós</em></td>
<td>2PL: <em>vosotros,-as</em></td>
</tr>
<tr>
<td>2PL: <em>vocês</em></td>
<td>2PL: <em>am a m</em></td>
</tr>
<tr>
<td>2PL: <em>os senhores,</em></td>
<td>2PL: <em>ustedes</em></td>
</tr>
<tr>
<td>3PL: <em>eles,</em></td>
<td>3PL: <em>éllos,</em></td>
</tr>
</tbody>
</table>

Duarte (1995) discusses briefly the influence of tense, verb form and transitivity on the expression of pronominal subjects in BP. In her data, transitive verbs have a higher expression rate (75%) than intransitive verbs (64%) or copulas (69%), but it is not clear why this should be the case. The expression of subject pronouns is more common in the imperfect and present than in the preterit of indicative (Duarte 1995:57). As for the influence of tense, Duarte’s (1995) data shows a decrease in subject pronoun presence in the preterit. In Posio’s (2008) data from first person singular in EP, there is a difference between imperfect, present and preterit (see Table 8).

Table 8. Comparison between Duarte’s (1995) and Posio’s (2008) data from subject expression in different tenses in BP and EP

<table>
<thead>
<tr>
<th></th>
<th>BP (Duarte 1995:57)</th>
<th>EP (Posio 2008:73)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imperfect</td>
<td>73%</td>
<td>70.2%</td>
</tr>
<tr>
<td>Present</td>
<td>74%</td>
<td>48.2%</td>
</tr>
<tr>
<td>Preterit</td>
<td>31%</td>
<td>39.2%</td>
</tr>
</tbody>
</table>

Looking at the percentages in Table 8, one might feel tempted to predict, for example, that the major differences between BP and EP are to be found in the present tense and
not in the imperfect or preterit. However, not having access to the raw frequencies in Duarte’s data, a meaningful quantitative comparison between BP and EP is not possible based on the percentages given in Duarte (1995) and Posio (2008). Contrastive studies of BP and EP are certainly needed in order to fully understand the differences between the two varieties.

Silveira’s (2011) recent dissertation studies the connection of the semantic class of the predicate with subject pronoun expression in BP. According to Silveira’s (2011) analysis, the rate of subject pronoun expression depends on verb semantics in first and second person singular, whereas in third person singular other factors (such as discourse continuity) are more important and verb semantics do not appear to have a significant effect. In general, the factors affecting the expression of pronominal subjects vary from one person to another, and Silveira (2011:178) thus stresses the importance of analyzing different grammatical person separately rather than focusing on the factors that affect subject expression in all persons. Similarly to the present work, Silveira (2011) also notes that the most frequently occurring subject-verb combinations tend to occur in constructions where the presence or absence of the subject pronoun may be relatively fixed (see Posio 2011, 2012a, 2012b, to appear).

Interestingly enough, the effect of the semantic verb class is mostly due to the most frequent lexemes in each class, viz. dizer ‘say’, achar ‘think’, saber ‘know’ and ter ‘have’ (Silveira 2011: 119). If the most frequently occurring verbs (defined as those constituting at least 1% of the data) are removed from the analysis, semantic verb classes do not show a significant effect (Silveira 2011:169; cf. Erkel & Guy, to appear). The distinct patterns of subject pronoun expression or omission observed with the most frequently occurring verbs are attributed to frequency effects. For instance, Silveira (2011: 185) argues that the low frequency of subject expression associated with saber ‘know’ is due to the conserving effect of high frequency (see Bybee & Thompson 1997, Bybee 2010). The most frequently occurring contexts of subject pronoun usage are argued to resist the ongoing increase in the rate of subject pronoun expression in BP.29

All in all, Silveira’s (2011) study covers similar grounds as the present dissertation, although in the current work the main focus is on the comparison of EP and PS. It is interesting to note that despite the differences in overall frequency of subject pronoun expression in the Brazilian and European varieties of Portuguese, verb semantics, frequency and the development of formulaic sequences or constructions play a crucial role in both varieties, at least as far as first person singular is concerned. Comparing BP and EP data using the same methodological and theoretical approach would be needed in future work to assess the importance of frequency effects in both varieties.

---

29 It is not clear, though, why other frequently occurring first person singular verbs rather favor than disfavor subject pronoun expression (cf. Silveira 2011: 184). See Section 5.4 for a discussion.
3.2.2. Differences between Brazilian and European Portuguese and Spanish

Duarte (1995:7-17) studies subject expression in BP but also presents quantitative results from EP, using data drawn from a corpus of sociolinguistic interviews. Table 9 (repeating, for convenience, the EP data presented in Table 2) shows the frequency of subject expression in different grammatical persons in Duarte (to appear). 30

In Duarte (2000:25), slightly lower percentages of subject expression are cited for EP: 35% for first person, 24% for second person and 21% for third person (the raw frequencies are not given). Barbosa, Duarte & Kato (2005:23) report a lower frequency of expressed pronominal subjects and NP subjects in the third person, viz. 22%, but their data consists of written journalistic interviews and contains only 162 clauses. None of these studies analyzes the EP data in order to search for differences between tenses, moods, verb lexemes or verb types, as is frequently done in the research on Spanish.


<table>
<thead>
<tr>
<th></th>
<th>European Portuguese (Duarte, to appear)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1SG</td>
<td>41% (173/421)</td>
</tr>
<tr>
<td>2SG</td>
<td>39% (54/140)</td>
</tr>
<tr>
<td>3SG</td>
<td>25% (28/110)</td>
</tr>
<tr>
<td>1PL</td>
<td>32% (9/28)</td>
</tr>
<tr>
<td>2PL</td>
<td>28% (80/285)</td>
</tr>
<tr>
<td>3PL</td>
<td>26% (34/132)</td>
</tr>
</tbody>
</table>

According to Duarte (1995:17), null subjects are preferred in EP in all clause types except in relative clauses, indirect interrogatives and interrogatives with question words. However, since the author does not present the raw numbers but only percentages, it is impossible to estimate the statistical significance of the findings. The age or level of instruction of the speakers was not found to be relevant to the frequency of expressed subjects (Duarte 1995). According to Barbosa, Duarte & Kato (2001:544), BP and EP differ from each other in that when the antecedent of a subject is not found in the immediately adjacent clause, the expression of pronominal third person subjects is “free” in EP but obligatory in BP. Barbosa, Duarte & Kato (2005:47) come to the conclusion that, in both BP and EP, “the null subject is favored under the condition of topic maintenance, while the overt variant is favored when there is a topic shift”. Concerning the differences between BP and EP, Duarte (2000:27) concludes the following:

On the basis of the analysis so far, it seems that, in EP, filling the subject position does not seem to be an option in embedded contexts or in those where there is a well-defined topic; a null category is the unmarked form of the subject.

30 Duarte’s original study focusing on European Portuguese is cited in Duarte (1995) as “in press”. It remains unpublished at the time of writing the present paper (Maria Eugênia Duarte, p.c., March 23rd 2011), which is why I refer rather to the 1995 paper in the following discussion. However, the data shown in Tables 2 and 9 are cited from Duarte’s (to appear) unpublished manuscript because the raw frequencies are given there but omitted from Duarte (1995).
In this respect, subject expression in EP does not appear to differ from Peninsular Spanish. As for the influence of verb lexeme, verb type or the pragmatic functions of the presence or absence of pronominal subjects, there is no data from EP presented in previous research.

A further difference claimed to exist between the two varieties of Portuguese is the possibility to refer to inanimate entities by pronominal subjects. In BP third person personal pronouns are used for inanimate referents quite frequently, while according to Barbosa, Duarte & Kato (2005:23) “overt subject pronouns in EP are almost invariably [+animate]”. In Barbosa, Duarte & Kato (2001:544) it is even claimed that null subjects are used for inanimate referents in 100% of the PE data but in only 57% of the BP data. Such generalizations are too strong, considering that they are based on a reduced number of examples consisting only of written interviews published in a Portuguese newspaper. 31 In Duarte’s data from spoken EP, pronominal subjects are used instead of null subjects to refer to as much as 7% of the inanimate referents (Duarte 2000:15). In practice it is not difficult to find examples of pronominal subjects with inanimate referents in EP. Thus consider examples (7)–(9).

(7) [...] mas um bocadinho de **broa** quando ela é assim
but ART bit-DIM of corn.bread when she be-3SG like.this

*a little bit of corn bread when it is like this*

ainda é melhor que o trigo [...] even be-3SG better than ART wheat

(it) is even better than wheat…” (CdP, Cordial: COV07)

(8) (speaking of a tree variety:)

Mas ela não tem folhas brancas? [...] Toda ela é amarela?
but she NEG have-3SG leaves white [...] all she be-3SG yellow

*But doesn’t it have white leaves? It’s all yellow?’ (CdP, Cordial: MIG29)

(9) **Ele** é o **Estado**... O **Estado**, ele pode pagar aos quinhentos
he be-3SG ART state ART state he can-3SG pay-INF PREP:ART 500

*It’s the state. The state, it can pay 500 escudos per year go-3SG.SUBJ that he pay-3SG.SUBJ PREP 1500 escudos.

escudos por ano... Vá que ele pague a um conto e quinhentos.
escudos per year. Why doesn’t it pay 1500 escudos?’ (CdP, Cordial: ALV03)

While the proportion of pronominal subjects referring to inanimate referents intuitively seems to be lower in EP than in BP, it is not a feature categorically distinguishing the two varieties. Rather, there is a difference between both varieties of Portuguese and Spanish (at least the Peninsular variety) where the use of pronominal subjects for inanimate referents is extremely rare: in Enríquez’s (1984:177) data, only 0.15% of the expressed pronominal subjects had a non-human referent. The use of pronominal subjects for inanimate entities in EP requires further research. Although the question is not central to the present work, it supports the hypothesis that EP differs from PS with respect to the characteristics generally attributed to null subject languages, such as the exclusive use of null subjects for non-human referents.

31 As a matter of fact, the EP corpus contains only 36 overt subjects out of which only one has an inanimate referent (Barbosa, Duarte & Kato 2005:23).
4. Data and methodology

The comparability of data drawn from two different languages may be problematic from two perspectives: the comparability of the languages in general and the comparability of the particular data representing these languages. The first question is addressed in Section 4.1 and the second in Section 4.2. Section 4.3 discusses the methodology adopted in the current research.

4.1. Typological profile of Peninsular Spanish and European Portuguese

Peninsular Spanish and European Portuguese are closely related Ibero-Romance language varieties. Given their close genealogical and areal proximity, it is not surprising that their typological profiles are very similar. The World Atlas of Language Structures (Dryer & Haspelmath 2011; henceforth referred to as WALS) has in total 58 features with a value provided for both Spanish and Portuguese,\(^{32}\) and in only nine cases are the languages reported to have a different value. The features with different values for Spanish and Portuguese are listed in Table 10.

In most of the nine factors listed in Table 10 it is actually questionable whether the different values reflect genuine differences between Portuguese and Spanish or whether they are due to different interpretations of the source literature. Some of the interpretations may even be considered erroneous. For instance, the Portuguese tense system does include an inflectioal future tense, cognate with the Spanish inflectioal future (cf. Feature 67A). There is also a morphological imperative for second person plural in Portuguese: it is obsolete in most varieties, but the same is true of most varieties of Spanish as well (cf. Feature 70A).\(^{33}\) To the best of my knowledge, out of the features listed in Table 10 the only one where Portuguese and Spanish clearly differ from each other is the word for ‘tea’ (Feature 138A), which is unlikely to be of any relevance for the present study. Thus PS and EP are indeed not only genealogically and geographically but also typologically very close to each other, and one might indeed to expect to find little differences in the use of pronominal subjects between the two languages.

\(^{32}\) In WALS, no distinction is made between Brazilian and European varieties of Portuguese or different areal varieties of Spanish. This might be problematic if the varieties are as differentiated from each other as Brazilian and European Portuguese, as it is not always clear which of the variants is chosen as representative. On the other hand, it is clear that a quantitative typological survey of the magnitude of WALS cannot be expected to cover any language-internal variation.

\(^{33}\) It is also unclear to me what is meant by “initial polar question particles” in Portuguese. Portuguese has the question particle *é que*, but it is neither fixed to the initial position nor used in polar questions.
Table 10. WALS features with different values for Portuguese and Spanish

<table>
<thead>
<tr>
<th>Feature</th>
<th>Portuguese</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>16A: Weight Factors in Weight-Sensitive Stress Systems</td>
<td>Lexical stress</td>
<td>Combined</td>
</tr>
<tr>
<td>46A: Indefinite Pronouns</td>
<td>Mixed</td>
<td>Special</td>
</tr>
<tr>
<td>67A: The Future Tense</td>
<td>No inflectional future</td>
<td>Inflectional future exists</td>
</tr>
<tr>
<td>68A: The Perfect</td>
<td>No perfect</td>
<td>From possessive</td>
</tr>
<tr>
<td>70A: The Morphological Imperative</td>
<td>Second singular</td>
<td>Second singular and second plural</td>
</tr>
<tr>
<td>82A: Order of Subject and Verb</td>
<td>SV</td>
<td>No dominant order</td>
</tr>
<tr>
<td>92A: Position of Polar Question Particles</td>
<td>Initial</td>
<td>No question particle</td>
</tr>
<tr>
<td>116A: Polar Questions</td>
<td>Initial question particle</td>
<td>Interrogative word order</td>
</tr>
<tr>
<td>138A: Tea</td>
<td>Words derived from Sinitic cha</td>
<td>Words derived from Min Nan Chinese te</td>
</tr>
</tbody>
</table>

The values related to the expression of pronominal subjects and word order are of special interest for the present discussion (see Table 11). In WALS, both Spanish and Portuguese are classified as languages with subject affixes on verb. In both languages, the basic word order in transitive clauses is SVO and the order of object and verb is VO. However, Spanish and Portuguese are claimed to differ regarding the order of subject and verb in intransitive clauses: Spanish is said to have “no dominant order” whereas Portuguese is reported to display SV order. Although this difference is far from categorical in practice, it is in line with the observation that postverbal subjects are “rare” in Portuguese (Ambar 1992; Posio 2008:37, 2012a). In general, flexibility of constituent order has been argued to co-occur with optionality of subject expression (Chomsky 1981; Gilligan 1988; cf. Roberts 1993 on French and Morales 1997 on Caribbean Spanish).

Table 11. WALS features concerning the expression and placement of subjects

<table>
<thead>
<tr>
<th>Feature</th>
<th>Portuguese</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>81A: Order of Subject, Object and Verb</td>
<td>SVO</td>
<td>SVO</td>
</tr>
<tr>
<td>82A: Order of Subject and Verb</td>
<td>SV</td>
<td>No dominant order</td>
</tr>
<tr>
<td>83A: Order of Object and Verb</td>
<td>VO</td>
<td>VO</td>
</tr>
<tr>
<td>101A: Expression of Pronominal Subjects</td>
<td>Subject affixes on verb</td>
<td>Subject affixes on verb</td>
</tr>
</tbody>
</table>

Given the typological proximity of PS and EP, the differences found in the actual tendencies of subject pronoun expression between the two languages suggest that typological proximity actually tells little about the expression of pronominal subjects. On the other hand, it should be borne in mind that typological parameters such as the ones used in WALS do not capture statistical differences manifested in language usage but rather describe prototypical instances. This problem has been acknowledged in the domain of linguistic typology, and there have been moves from traditional feature-based classifications towards more statistically oriented approaches that use comparable data such as different language versions of the Pear Stories (Bickel 2003) or translated
parallel corpora (Cysouw & Wälchli 2007) instead of relying on reference grammars and other secondary sources.

Although the typological import of the present study remains modest given that it examines only two closely related languages, the theoretical significance of the findings lies precisely in showing that despite being classified as typologically similar, the languages under survey manifest significant statistical differences with regard to a specific typological feature. If such differences were observed between unrelated, typologically distinct languages, they would naturally be less interesting and could be simply attributed to the trivial observation that languages tend to be different from each other.

In order to compare two or more different languages, the comparisons must generally be based on the use of comparative concepts (Haspelmath 2010), not on language-specific morphosyntactic categories. For instance, there is no point in comparing “the dative case” in two different languages as cases are typically polysemous and the meanings conveyed by a case with the same name in two languages typically overlap only partially: it is more meaningful to base the comparison in semantic components that are judged universal, such as the semantic role Beneficiary. This principle is implemented in the present study through the use of semantic roles instead of categories based on verb lexemes in order to operationalize the impact of semantic factors (see Posio to appear). Although such notions as “subject” or “present of indicative” are not comparative concepts in the universal sense, their use in the present study is justified, given that the categories are sufficiently similar in the two languages: first person singular subjects can be recognized formally by verbal morphology and semantically by reference to the speaker in both languages.

One of the factors that permit the comparison of subject pronoun usage in PS and EP is the similarity of their person marking paradigms (see Table 7). The choice of first person singular as the main object of analysis is justified by the fact that it is the most comparable grammatical person in the two languages. It bears distinctive, non-syncretic person marking in the present, preterit and future of indicative in both languages. In the imperfect of indicative, conditional and all subjunctive forms, the verb form is syncretic with third person singular (e.g. EP amava, PS amaba ‘I/(s)he loved’). Thus any eventual differences found between EP and PS cannot be explained away by syncretic person marking in one of the languages (see Section 3.2.1).

Table 12 summarizes the person syncretisms found in Spanish and Portuguese.

<table>
<thead>
<tr>
<th>Type of syncretism</th>
<th>Tenses and moods affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1SG = 3SG</td>
<td>Imperfect of indicative</td>
</tr>
<tr>
<td></td>
<td>Conditional</td>
</tr>
<tr>
<td></td>
<td>All subjunctive tenses</td>
</tr>
<tr>
<td>2PL = 3PL</td>
<td>All tenses and moods in Portuguese</td>
</tr>
<tr>
<td></td>
<td>All tenses and moods in Andalucian, Canarian and American varieties of Spanish</td>
</tr>
</tbody>
</table>

In the second person singular, EP has a tripartite distinction between the familiar address form tu, semi-familiar você and deferential or formal o senhor (masculine) and
*a senhora* (feminine), whereas PS has only two address forms, the more formal *usted* and the familiar *tú*. In addition, while PS seems to favor the overt expression of *usted*, presumably as a politeness strategy (cf. Enríquez 1984), in EP expressing the pronominal address form is generally avoided. This makes comparing second person pronouns a complicated task that would require careful examination of such sociolinguistic and sociosituational information that is not provided in the corpora examined.

### 4.2. The data and their comparability

The data analysed in the current work are drawn from freely available corpora of spoken Peninsular Spanish and European Portuguese. These corpora were chosen because they were the most accessible and most representative ones available at the beginning of the dissertation project. The corpora used are listed below:


3. **CREA** = REAL ACADEMIA ESPAÑOLA. *Corpus de referencia del español actual.*

4. **CdE** = Davies, Mark. (2002-) *Corpus del Español* (100 million words, 1200s-1900s).

5. **CdP** = Davies, Mark and Michael Ferreira. (2006-) *Corpus do Português* (45 million words, 1300s-1900s).

6. **CORDIAL-SIN** = Syntax-oriented Corpus of Portuguese Dialects. Linguistics Center of Lisbon University.

Given that the corpora represent two different languages and they have been collected for different purposes, the criteria of data collection and text selection are not exactly the same for both languages. The EP resources consist mostly of interviews or elicited conversations, while for PS there is a larger selection of texts available, especially in COREC and CREA which are organized into various subcorpora according to the medium and type of situation where the data was collected. The CORDIAL-SIN contains dialectal interviews which are a genre of their own as regards the variety of topics discussed and the type of interaction between the speakers. The Corpus del Español and the Corpus do Português are compilations of data drawn from smaller corpora.

Using ready-made corpora is naturally not an ideal starting point for a comparative study: ideally, the data should be gathered applying similar data collection and transcription methods and assuring that the data come from similar communicative situations for both of the languages under survey. However, since the methods used in
the present study call for relatively large amounts of data, collecting and transcribing all data by myself would have been an impossible task in the scope of a four-year dissertation project. The use of parallel (i.e. translation) corpora (cf. Seo 2001) is also out of the question given that the study focuses on spoken language. Therefore I have tried to make the best possible use of the existing corpus resources. Table 13 lists the use of data in the articles.

Table 13. Use of PS and EP speech corpora in the articles

<table>
<thead>
<tr>
<th>Article</th>
<th>PS data</th>
<th>EP data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Posio 2011. Spanish subject pronoun usage and verb semantics revisited: first and second person singular subject pronouns and focusing of attention in spoken Peninsular Spanish</td>
<td>Random samples of 100 occurrences of 14 verbs in first and second person singular from the CREA (oral, Peninsular Spanish).</td>
<td>–</td>
</tr>
<tr>
<td>2. Posio 2012a. Functions of postverbal pronominal subjects in spoken Peninsular Spanish and European Portuguese</td>
<td>All clauses with a postverbal subject pronoun from the oral Peninsular Spanish data in the Corpus del Español.</td>
<td>All clauses with a postverbal subject pronoun from the oral European Portuguese data in the Corpus do Português.</td>
</tr>
<tr>
<td>3. Posio 2012b. Who are ‘we’ in spoken Peninsular Spanish and European Portuguese? Expression and reference of first person plural subject pronouns</td>
<td>A selection of 20 texts from the COREC.</td>
<td>The 20 texts collected in the 1990s representing European Portuguese in the Português Falado and one text from the CORDIAL-SIN.</td>
</tr>
</tbody>
</table>

In Posio (2011), the data is drawn from the oral subsection of Peninsular Spanish in CREA corpus and in Posio (2012a) the data consist of all clauses with postverbal pronominal subjects found in the Corpus del Español and Corpus do Português. Posio (to appear) and (2012b) compare PS and EP to each other and the same selection of data is used for both articles. For EP, this data consists of all texts gathered in the 1990s in Portugal included in Português Falado and one text from CORDIAL-SIN. In order to ensure the comparability of the data, the PS data were chosen from the subcorpora Entrevistas (‘interviews’) and Conversaciones (‘conversations’) of the COREC, taking care to include data from similar speech situations and communicative genres that are represented by the texts in Português Falado. Despite this effort, the PS selection contains both spontaneous conversations and more formal interviews (e.g. from TV and radio programs) while the EP data consist of conversations and interviews realized specifically for the purposes of the corpus collection. Despite these differences, the selections of data used in (Posio to appear) and (2012b) were nevertheless found to be quite similar with regard to the distribution of the semantic role of the subject, tenses and moods and same vs. switch reference subjects. This observed similarity can be taken as indication that the corpora are sufficiently similar to be comparable and that the differences found are not attributable to different genre effects.
The fact that large corpora such as Corpus del Español or Corpus do Português often contain a highly heterogeneous selection of texts and incorporate contents from smaller corpora, collected using different selection and annotation criteria, is problematic for all corpus-based research; care should be taken when defining the scope of generalizations that can be drawn from any particular set of data. In the current work, the aim is not to consider genre-specific, sociolinguistic or situational variation. In principle, no claims are made concerning particular genres such as interviews or spontaneous conversation: rather, the object of study is defined as spoken language as represented by the corpora. Examining eventual differences between different communicative genres must be left to future research.

4.3. Methods

The methodology used in the present study pertains to the domain of corpus-based linguistics where data extracted from corpora are used to support or falsify hypotheses (Tognini-Bonelli 2001). However, the hypotheses themselves are based on preliminary findings from corpora (Posio 2008, 2010) and new hypotheses and research questions emerge during the research process. Conducting a corpus-based research is a dialogical process where one starts with certain hypotheses and premises but is also required to modify or change the initial hypotheses in the course of the process.

In order to ascertain the reliability of the results, the present study relies on triangulation of the methodology and the data whenever possible. The principle of triangulation refers to testing the hypotheses using different data (i.e. data triangulation) and different types of methods (i.e. method triangulation) in order to ascertain the validity of the results (see Denzin 1970). In concrete terms, triangulation is manifested in following ways in the present study:

1. Different selections of data from different corpora are used in the four articles of the dissertation (see Table 13).

2. Different data collection methods are used in different studies, namely random sampling (Posio 2011), analyzing all clauses in a small-scale corpus (Posio 2012b, to appear) and selecting all relevant clauses in a larger-scale corpus for the analysis (Posio 2012a).

3. Different analytic approaches are combined in all studies. The articles present quantitative results and use the Pearson’s chi-square test and chi-square residuals to test the statistical significance of the findings (in particular Posio 2011 and to appear). However, examination of quantitative data such as frequency counts is constantly used along with qualitative scrutiny of representative examples in order to interpret and explain the quantitative findings.

Note, however, that the contents of the Portuguese corpora overlap partially, as Corpus do Português contains part of the texts included in CORDIAL-SIN and Português Falado. While this problem is fully acknowledged, it it cannot be completely avoided given the limited availability of speech corpora from EP.
The main focus of the current work is on pragmatics and semantics and the analysis is monofactorial. In other words, the intention is not to evaluate the relative importance of different factors but, rather, to deepen the understanding of the semantic and pragmatic factors that have shown to be relevant in previous research. The qualitative analysis consists of the examination of representative examples in their context of use. As for the smaller corpora used in Posio (to appear) and Posio (2012b), it was possible to access the whole transcript where the examples were taken from. In Posio (2011) and Posio (2012a), the window of observation was limited to the amount of context given by the search interface in the corpora, varying between 100 and 200 words. This was usually sufficient to observe the preceding and following turn in the discourse (in case of doubt, another example was selected for the qualitative analysis).

The characteristics of the data naturally place restrictions to the methods that can be used in the analysis and the type of observations to be made. Given that the studies are conducted using ready-made corpora relying on orthographic transcription, many interactionally relevant phenomena including overlapping speech, length of pauses, prosodic features and non-linguistic behaviour of the speakers (e.g. gaze and facial expressions) cannot be taken into account. Thus the qualitative analysis of examples is limited to examining those aspects of discourse that are transcribed in the corpus: this is inevitable in order to be able to use sufficiently large corpora of spoken language to permit the examination of frequency. Future work is needed in order to account for the factors left out of the scope of the present study.
5. Results and conclusions

Section 5.1 presents a brief summary of the contents of the articles of this dissertation. Sections 5.2–5.4 discuss the main results of the whole dissertation, and the main conclusions from all of the articles are drawn in Section 5.5.

5.1. Summary of the articles


The first article (Posio 2011) focuses on Peninsular Spanish data in the interest of revisiting the question of how verb semantics and subject pronoun expression may be connected to each other in Spanish. In most previous studies, verb semantics and grammatical person have been regarded as two factors potentially affecting the expression of pronominal subjects, but less effort has been put into explaining what the connection between verb semantics and subject pronoun expression is. The analysis focuses on first person singular, although data from second person singular is also analyzed and contrasted to first person singular data. For both persons, the data is drawn by random sampling from 14 frequently occurring verb lexemes in a corpus of spoken PS (i.e. the oral section of CREA), using only clauses with present of indicative verb forms in order to minimize the influence of other factors such as switch reference or tense/mood/aspect of the verb to the results. The expression of first and second person singular subjects is shown to be less frequent in highly transitive clauses than in poorly transitive ones. In highly transitive clauses, attention is hypothesized to be focused more on the Patient or on the action expressed by the verb than on the Agent. In this article, it is suggested that subject pronoun usage is partially motivated by event semantics and the relative salience of the different participants of the event, although the importance of formulaic sequences with fixed subject pronoun usage is also recognized.


The second article (Posio 2012a) moves the analysis from semantics of the event and the verb to pragmatic aspects of subject pronoun usage. The data analyzed are drawn from relatively large corpora of spoken language (Corpus del Español and Corpus do Português) and consist of only those clauses where a pronominal subject is placed after

---

35 The verb lexemes were chosen among the most frequently used verbs in Spanish, leaving out auxiliaries (e.g. haber ‘have’), modals (e.g. poder ‘can’) and the verb saber ‘know’ due to its frequent occurrence in formulaic sequences (which were not among the main concerns of the dissertation at this stage of the work).

36 The corpora and the abbreviations used of them are presented in Section 4.2.
the verb. This is the only article of the dissertation that analyzes all grammatical persons: this choice is justified by the lack of previous empirical, quantitative surveys on postverbal pronominal subjects in PS and EP. Although the variation between subject-verb and verb-subject order in both Spanish and Portuguese has been studied rather extensively, the previous studies have either not considered pronominal subjects at all or have not distinguished between pronominal subjects and noun phrases. This article shows that the factors motivating verb-subject order with pronominal subjects differ fundamentally from what has been claimed to motivate the postverbal placement of noun phrases. Information structure (in concrete, the distinction between more and less accessible referents or given and new information) is less relevant for pronouns, as they practically always represent given information in discourse. Other factors argued to favor postverbal placement of subjects, such as unaccusativity or intransitivity of the verb, are not relevant either. Rather, in addition to the contrastive use, in PS the postverbal placement of subject pronouns often serves the function of backgrounding the subject or foregrounding another element in the clause. Although subject pronoun expression in the preverbal position is more frequent in EP than in PS, the postposition of subject pronouns is very rare in the EP data and most of the cases are found in formulaic sequences such as quotatives or epistemic parentheticals. In the PS data, postposed subject pronouns are more frequent than in EP and tend to occur more productively, i.e. outside of discourse formulae. In comparison with the first article (Posio 2011), the notion of formulaicity gains more prominence in this article, as it is shown that especially in EP postverbal pronominal subjects are found predominantly in formulaic sequences.


The third article (Posio 2012b) looks at subject pronoun expression in first person plural. First person singular is not included in the analysis, but the comparison between first person singular and plural is possible because the corpora examined are the same as in Posio (to appear). On the one hand, this study provides evidence for the hypothesis that subject pronoun usage is motivated by different factors in different persons (see Section 1.1). On the other hand, it is shown that PS and EP differ significantly with regard to the frequency and motivations of subject pronoun expression in first person plural (1PL). In the PS data, the expression of 1PL subject pronouns is very rare (4.5%) compared to the EP data (where 32.2% of the 1PL subjects are expressed). This is connected to the observation that the 1PL subject pronouns tend to manifest different referential properties in the two languages. In PS the expressed subject pronoun nosotros implies a hearer-exclusive reading and is typically used to refer to a socially or professionally defined group of people construed as the agent of a joint action. In EP, the expression of the subject pronoun nós does not imply such an interpretation: it is compatible with hearer-inclusive, hearer-exclusive and impersonal readings. Thus the 1PL subject pronoun is used more frequently in EP than in PS. In addition, the article examines the difference between the older 1PL pronoun nós and the more recently grammaticalized pronoun a gente in EP, suggesting that the distribution of the two
pronouns is partially connected to verb lexemes. As for the 1PL pronouns proper (nosotros in PS, nós in EP), the presence or absence of the pronoun does not correlate with the verb or the semantic role of the subject like in first person singular.


The fourth article (Posio to appear) examines the ways in which semantic roles are defined in the literature and discusses their application to the quantitative analysis of authentic data. Relying on a frame-based approach to semantic roles, the analysis carried out in the article looks at the frequency of 1SG subject pronoun expression in different semantic roles in PS and EP. The corpora are relatively small and consist of a selection of oral texts from the corpora COREC, Português Falado and CORDIAL-SIN, and all clauses with 1SG verb forms are analyzed, viz. slightly over 700 clauses per language. It is shown that PS and EP differ significantly from each other with regard to the overall frequency of subject expression. In both languages, there are significant differences between different semantic roles as well. In the second part of the analysis, the data are examined from a more qualitative perspective, looking for potential motivations of especially high or low frequency of subject expression in connection to particular semantic roles. The analysis reveals that both PS and EP manifest a tendency towards less frequent subject expression in the Agent role and more frequent in the Stative role, but differ significantly from each other in the roles Cognizer and Communicator. In the latter roles, the differences can be attributed to different formulaic sequences with highly frequent verb lexemes, viz. ‘think’, ‘know’ and ‘say’. For instance, the verb form creo ‘I think’ is to a great extent responsible of the fact that the role Cognizer has a relatively high frequency of expressed subjects (compared to a relatively low frequency in EP). In EP, the highest frequency of expressed subjects is found in the Communicator role due to the tendency of expressing the subject in quotative constructions with the verb form digo ‘I say’; in PS, the Communicator role disfavors subject expression. As a conclusion, it is suggested that any analysis of authentic data based on semantic roles should be informed of the influence of formulaicity, pragmatic functions of formulaic sequences, and frequency of different items in speech. Analyzing spoken language relying only on semantic roles may lead to distorted results. In a contrastive perspective, this article shows that the most frequently occurring verb lexemes are largely responsible for the differences between the two languages in subject expression.

5.2. Peninsular Spanish and European Portuguese in comparison

One of the main descriptive goals of the present work is to provide new information on the use of pronominal subjects in PS and EP and on the differences between the two languages. As the results of Posio (2011, 2012b, to appear) demonstrate, the use of first person pronominal subjects (both singular and plural) is significantly more frequent in
EP than in PS in the preverbal position. The frequency difference between PS and EP is especially striking in the case of first person plural (Posio 2012b). As for the postverbal position, the expression of pronominal subjects is more frequent and more productive in PS than in EP (Posio 2012a). One might have suspected that the higher frequency of subject expression in EP would cause subject pronoun usage to be less sensitive to semantic and pragmatic factors (i.e. less “marked”) than it is in PS. However, in the case of first person singular, the differences in subject pronoun usage between different semantic roles turned out to be significant in both PS and EP, the difference being that in most contexts the overall frequency of expressed subject pronouns is higher in EP than in PS. Thus the idea of subject pronoun usage being more semantically or pragmatically marked in one language than in the other becomes indefensible.

Although PS and EP differ from each other with respect to the general subject expression rate, the tendencies related to the semantic roles of the subject in first person singular are basically very similar in both languages. This finding is parallel with those reported in previous research comparing different varieties of Spanish: although the varieties have significantly distinct rates of subject expression, the factors affecting subject expression (e.g. grammatical person, verb type, and coreferentiality) show nearly identical patterning across the varieties (see Cameron 1993; Travis 2007; Silva-Corvalán 2001; Torres Cacoullos and Travis 2011:250). The major differences between PS and EP, then, are found in those semantic roles where subject expression with determined verb forms has reached a high grade of formulaicity due to high frequency of use, namely in the roles Cognizer and Communicator that are frequently instantiated by the subjects of formulaic sequences such as quotatives or epistemic parentheticals.

Previous research has typically concentrated on explaining only the presence of subject pronouns and considered their omission as the default or unmarked option, given that Spanish and Portuguese are null subject languages. Starting from the observation that in certain contexts subject expression is equally or more frequent than omission in both PS and EP, the present research rather highlights the fact that both presence and absence of pronominal subjects have their own pragmatically motivated functions in different contexts and none of the two options can be considered as default. An example of this is the construction (yo) creo que in PS and (eu) echo que in EP (both of which can be translated as ‘I think that’) where the presence or absence of the pronominal subject is related to the level of speaker commitment to the proposition being modified by the constructions (see Posio 2011, to appear).

Previous empirical research on Romance subject pronoun usage has often paid less attention to the postverbal placement of pronominal subjects, considering it as marginal or merging preverbal and postverbal pronominal subjects into same category. Posio (2012a) concentrates precisely on this area of subject pronoun expression and demonstrates that postverbal subject pronouns are used more frequently in PS than in EP in all grammatical persons. In EP, the postverbal placement of pronominal subjects is less productive and occurs mostly in restricted, formulaic uses with specific verb lexemes such as dizer ‘say’ in quotatives. The factors that have been argued to motivate the postverbal placement of subjects in general (such as verb type or informational status of the subject referent) are mostly irrelevant in the case of pronominal subjects.

What is more, in addition to the contrastive uses highlighted in previous research, the
postverbal position also serves to background the subject or the stretch of discourse where it occurs, signaling reduced agentivity or topicality of the subject.

The differences found in the frequency of subject pronoun expression between PS and EP cannot be explained away by syncretic person marking. Although EP person marking system is affected by more syncretisms than PS (see Section 3.2.1 and Table 7), none of the syncretisms affects first person singular or plural more in one of the languages than in the other. Attributing the more frequent expression of subject pronouns in EP to genre differences does not seem plausible either (see discussion in Posio 2012b), as the data from both languages represent similar genres (interviews and conversations). In addition, the fact that PS data has a higher proportion of conversational texts should rather increase the frequency of subject pronoun expression than decrease it (cf. Travis 2007). Even if not using a subject pronoun in such contexts may be the most frequent option, it is not to be considered neutral with respect to communicative functions.

The four studies included in this dissertation show that despite being genealogically, typologically and geographically close to each other, EP and PS differ in that the expression of pronominal subjects is considerably more frequent in the former than in the latter, with the exception of the postverbal position (Posio 2012a). In a wider typological context, this observation challenges the traditional view of null subject languages as a uniform category where the expression of pronominal subjects is rare and reserved e.g. for contrastive or emphatic functions (Dixon 2010:117). Future comparative work is needed in order to fully understand or account for the differences in the distribution of pronominal subjects (see Seo 2001, Bickel 2003).

5.3. Subject expression, verbs and grammatical persons

It has long been known that subject pronouns are used at different rates in different grammatical persons (e.g. Cameron 1992, 1995, 1997; Duarte 1995; Enríquez 1984). The present study points out that the frequent use of subject pronouns in first person singular, in comparison to other persons, is related to the fact that a great proportion of first person singular verbs are used to express the speaker’s epistemic stance. It is not just that speakers like to speak of themselves (Davidson 1996: 553): they use first person singular verbs for different purposes than verbs in other persons (see Scheibman 2001). In this respect, first person singular differs notably from first person plural which does not serve stance-taking or other discourse purposes to the same extent as first person singular (Posio 2012b). A further difference between deictic persons (such as first person singular) and anaphoric persons (i.e. third person singular and plural) is that the latter need to be introduced into the discourse and represent different levels of accessibility or givenness to the discourse participants. While the use of null subjects is considered to require (or signal) a high level of accessibility (Ariel 2001), in deictic persons the role of accessibility is less clear, given that their referents are inherently accessible by virtue of being present in the communicative situation.

Chart 2. and Chart 3. show the distribution of these verb forms (both with and without expressed subject pronouns) in the data used in Posio (2012b, to appear). The verbs meaning ‘think’, ‘say’ and ‘know’ are among the most frequently used in both languages even though their ranking differs between PS and EP. In EP the verb dizer ‘say’ is less frequently used than its cognate decir in PS.37

Although the verbs ter ‘have’ and ir ‘go’ are more frequent than dizer ‘say’ in EP, they do not show a similar tendency towards formulaization and the entrenchment of subject pronoun usage or omission depending on the pragmatic meaning of the formula in question. This is because ter ‘have’ and ir ‘go’ are more general in meaning and occur in a wider variety of contexts, to the extent that both have grammaticalized as auxiliaries. The contexts of occurrence of the verb dizer ‘say’ are more limited: in the EP data, it is most frequently found in formulaic sequences with a quotative function (Posio 2012a, to appear).

37 In PS, the frequent use of decir is associated with formulaic uses of the verb in sequences functioning as discourse markers or quotatives; for details, see Posio (2011).
The fact that the verbs meaning ‘think’, ‘know’ and ‘say’ occur so frequently in speech has a two-way connection to their usage. On the one hand, their frequency is boosted by their usefulness in stance-taking and other discourse functions (Posio to appear). On the other hand, the frequent exposure of speakers to these particular forms in a limited set of collocations alters the way these forms are represented and accessed in speakers’ memory (see Bybee 2010, Travis & Torres Cacoullos 2012, Silveira 2011, Erkel & Guy to appear). The use of subject pronouns and other items that occur frequently with these verb forms is more likely to become entrenched and associated with the particular pragmatic function and context where the form occurs than the use of less frequently occurring forms (cf. Erkel & Guy to appear, Silveira 2011).

The most frequently used verbs are characterized by high token frequency but low type frequency: they occur frequently in a restricted number of collocations but are less productive in terms of tense, mood and aspect variation and the contexts of occurrence. A comparison between first person singular and plural provides further evidence of the connection of high token frequency and the development of formulaic sequences: none of the first person plural forms investigated occur with sufficient frequency in the data to give rise to formulaic sequences comparable to those found in first person singular (Posio 2012b). Thus verb lexeme, in addition to same/switch reference, is a factor affecting subject pronoun expression in first person singular in both PS and EP.

In studies focusing on frequency effects, it is sometimes assumed that high frequency alone may suffice to explain an item’s deviant syntactic behavior. For instance, Silveira (2011:185) considers that the low frequency of subject pronoun expression associated with the verb saber ‘know’ in Brazilian Portuguese can be attributed to the conserving effect of high frequency. As subject pronoun usage is demonstrably becoming more frequent in Brazilian Portuguese, the high-frequency item saber, especially in the first person singular collocation não sei ‘I don’t know’, is argued to resist the new tendency by virtue of being a frequently occurring item whose representation in the speakers’ memory has become strengthened due to frequent usage. However, it should be noticed that the verb saber ‘know’ is associated with low subject pronoun frequency also in PS and EP although there is no indication that the rate of subject pronoun expression would be increasing in other contexts in these languages.

In studies focusing on frequency effects, it is crucial to take into account the semantics of the construction and the pragmatic uses it serves in discourse. In Posio (to appear) it is remarked that the subject is typically not expressed in the formulaic sequence used to express the speaker’s epistemic stance, viz. PS não sé and EP não sei ‘I don’t know’, while the subject pronoun occurs more often in the non-formulaic use (cf. Scheibmann 2001). As the pragmatic function of the sequence is to reduce the speaker’s responsibility for the truth value of the utterance, the non-expression of the subject pronoun serves to avoid the focusing of attention to the speaker. Thus while the high frequency of the construction certainly is what makes possible the formulaization of the sequence and may lead to less frequent subject pronoun expression, it does not suffice to explain why the subject pronouns are left unexpressed in this construction in the first place. Although frequency plays a role in the entrenchment of formulaic sequences, the pragmatic functions of the sequences (and the semantic properties of the depicted events and their participants, cf. Posio 2011, to appear) are crucial in order to understand why certain constructions are more frequent than other.
Comparing subject pronoun expression in first person singular and plural (Posio 2012b) completes the picture of the development of formulaic sequences. Unlike in first person singular, the most frequent verb lexemes in first person plural are, in both PS and EP, ‘have’, ‘go’ and ‘be’ (see Posio 2012b for details). These verbs are general in meaning and highly productive, i.e. they occur in a wide variety of contexts, including auxiliary uses. Due to the lower frequency of first person plural and the lack of stance-taking or pragmatic uses typical of first person singular, the use or omission of subject pronouns with first person plural verbs shows no signs of entrenchment comparable to first person singular. The only noticeable exception is the systematic subject omission in the cohortative construction consisting of the verb form *vamos* (go-1PL) and a main verb in both PS and EP (Posio 2012b: 352). As observed in Posio (2012b), the use of subject pronouns is extremely rare in the PS data and tends to be reserved to hearer-exclusive uses. In the case of EP, the expression of first person plural pronominal subjects is sensitive to same/switch reference but fails to show particular lexeme-specific tendencies (see Posio 2012b for details).

5.4. Subject pronoun expression, semantics and pragmatics

One of the initial predictions of the present work (see Section 1.1) was that verb semantics are connected to the expression or omission of pronominal subjects in PS and EP. In Posio (2011, to appear) semantic roles of the subject are used as analytic tools permitting the comparison of semantics within two different languages. It is shown in Posio (to appear) that subject pronoun expression follows partially similar, partially different tendencies in PS and EP. Subject pronouns are expressed less often with Agent subjects than with Stative subjects in both languages (see Posio to appear for the definition of these semantic roles). As a semantic/pragmatic explanation, it is proposed that the expression of subject pronouns reflects a higher level of focusing of attention on the subject referent, whereas subject pronouns are left out more frequently when attention is focused on another participant or on the action expressed by the verb. Thus subject expression is more frequent with stative verbs than with agentive verbs or when there is an affected Patient in the clause (see Posio 2011, to appear). This trend can be observed in both PS and EP. However, the languages differ from each other in that the general expression rate of subject pronouns is higher in EP than in PS.

Semantic roles show a different effect in PS and EP in the case of the more specific roles, viz. Cognizer and Communicator (see Posio to appear). PS has a high rate of expressed subjects in the Cognizer while in EP this role does not differ from the general trend. On the other hand, while Communicator subjects are associated with a high rate of subject pronoun expression in EP, in PS subject pronoun expression is rare in this role. There is no evident functional explanation that could account for this difference between the two languages under survey. Thus it is crucial to observe which verb lexemes and forms occur in each semantic role. While Agent and Stative are fairly general roles that host a wide variety of verb lexemes, the roles Cognizer and Communicator are much more restricted in meaning and dominated by the most frequently occurring verb lexemes in the data (‘say’ in Communicator and ‘think’ in Cognizer). In Posio (to appear) it is argued that the high frequency of these particular
verb lexemes allows for different tendencies of subject pronoun expression to develop and entrench in the two languages under survey, causing PS and EP to differ most strikingly in these two roles. For instance, quotative constructions are associated with frequent subject pronoun usage in EP but rarely occur with subject pronouns in PS. In the case of Cognizer, the verb *creer* ‘think’ occurs frequently in the subjectivizing construction *yo creo que* ‘I think that’ in PS, causing the high frequency of subject expression in the Cognizer role. In EP, the distinction between this construction with and without subject pronoun is less clearly entrenched than in PS, and thus no distinctive pattern of subject pronoun expression emerges (Posio to appear).

A methodological conclusion that can be drawn from the results presented in Posio (to appear) is that although the use of semantic roles as analytic tools in cross-linguistic comparisons is useful and permits the comparison of categories that are not specific to any particular language (Haspelmath 2010), the expression of subject pronouns in semantic roles can be affected by frequency effects and formulaization (i.e. the development of formulaic sequences affecting frequently used items in language). Therefore it is important to take into account the frequency of particular lexemes and their pragmatic functions in languages use, not only the semantic roles they represent.

### 5.5. Conclusions

The present study has examined the use of pronominal subjects in first person singular and plural from a usage-based cognitive-functional perspective, using data drawn from speech corpora of Peninsular Spanish and European Portuguese. It has been shown that the two languages differ fundamentally from each other with regard to the frequency of subject pronoun expression and omission. In particular, it is demonstrated that frequency and formulaization are important factors in accounting for subject pronoun expression. It is argued that the omission of subject pronouns should not be considered a default option in null subject languages: rather, omission and expression have their own specific functions in different contexts of use. It follows from this that no single theory can fully account for subject pronoun usage in all persons and contexts: various explanations are needed.

The study also challenges the traditional assumption that subject expression in null subject languages (i.e. languages with variable subject expression) is used sparingly and reserved for the expression of contrast and emphasis. In addition, it is suggested that future typological studies based on frequency data from corpora may show that the category of “null subject languages” is actually more heterogeneous than has been thought. Considering that majority of the world’s languages belong to this group (Dryer 2011), it is perhaps surprising that cross-linguistic comparisons based on frequency data have been relatively rare among previous research.

The present study is, in many respects, limited to the available data and many factors that are potentially relevant in accounting for the use of subject pronouns have not been considered. These include the influence of informational status in the expression of anaphoric subjects, different types of priming effects, sociolinguistic factors, clause type, phonological and phonetic factors and intonation. In addition, the nature of the corpus data analysed does not permit a full-fledged examination of the
interaction between the speakers, including their non-linguistic behaviour in the communicative situation. All these must be left to future research. However, one of the contributions of the present work is that it presents descriptive, empirical results from European Portuguese and compares them with Peninsular Spanish. All of the articles open an array of potential future research topics. In particular, future research should focus on the comparison of different registers and text types, Brazilian and European (and other varieties) of Portuguese.
References


Claes, Jeroen. 2011.¿Constituyen las Antillas y el Caribe continental una sola zona dialectal? Datos de la variable expresión del sujeto pronominal en San Juan de Puerto Rico y Barranquilla, Colombia. Spanish in context 8:2, 191–212.


51


**Corpora**


