Brigham Young University
BYU ScholarsArchive

All Theses and Dissertations

2013-07-03

Pro-Drop and Word-Order Variation in Brazilian Portuguese: A Corpus Study

Stewart Daniel Smith
Brigham Young University - Provo

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Linguistics Commons

BYU ScholarsArchive Citation

Smith, Stewart Daniel, "Pro-Drop and Word-Order Variation in Brazilian Portuguese: A Corpus Study" (2013). *All Theses and Dissertations*. 3719.
https://scholarsarchive.byu.edu/etd/3719

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu.

Pro-Drop and Word-Order Variation in Brazilian Portuguese:

A Corpus Study

S. Daniel Smith

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Heather Willson Sturman, Chair
Mark Davies
Dan P. Dewey

Department of Linguistics and English Language

Brigham Young University

July 2013

Copyright © 2013 S. Daniel Smith

All Rights Reserved

ABSTRACT

Pro-Drop and Word-Order Variation in Brazilian Portuguese:
A Corpus Study

S. Daniel Smith
Department of Linguistics and English Language, BYU
Master of Arts

The present study examines cetain syntactic properties of the Brazilian variety of Portuguese (BP): 1) BP is a pro-drop language with instances of both null subjects and covert objects, and 2) BP exhibits several possible word orders. To determine the frequency of pro-drop and word-order variations, *the CDP* (The Portuguese Corpus) was used to provide samples of transitive, main clauses, which were then categorized based on whether or not they had null subjects and covert objects. The clauses were also categorized according to word order. In addition to providing samples, the corpus allowed for the comparison of four different registers of BP: academic, newspaper, fiction, and oral. The results of the present study demonstrated that null subjects are much more common than covert objects (29.4% and 2.3% respectively) and that register did significantly affect the frequency of pro-drop, with oral having the highest rate of pro-drop and newspaper the lowest. For word order, SVO was most common at 95.1% with the occurrences of other variations being too rare to reliably determine statistical significance. Different from pro-drop, register did not affect the frequency of different word orders. Word-order variations were not random, however, but were determined by *topic* and *focus* with old information (topic) generally occurring preverbally, and new information (focus) generally occurring in the most embedded position. The fact that this study effectively examined these syntactic features is significant, as most of the Portuguese syntactic research previous to the present study was specific to European Portuguese. The present study demonstrated a new methodology being successfully applied to a different dialect, but more than that, it demonstrated that a more empirical, data-driven approach to syntactic research is both possible and valuable, justifying the creation and use of large corpora for this type of research.

Keywords: [Brazilian Portuguese, pro-drop, null subject, word order, topic, focus, corpus]

ACKNOWLEDGEMENTS

There are many individuals without whom I could never have completed this study. I wish to specifically thank the members of my committee: Dr. Dan P. Dewey for his help with the statistical analysis, Dr. Mark Davies for assisting in the methodology design and for creating *the CDP*, and Dr. Heather Willson Sturman for her guidance, patience, and willingness to spend time on this project at a very busy time for her.

Eu sou muito grato também por Carlos, Rosa, Fernando e o resto da família dos Santos. Sem a paciência, ajuda e apoio deles, eu não teria conseguido completar este projeto. Não há palavras suficientes para contar nem para agradecer tudo que fizeram por mim, então só posso dizer obrigado.

Finally, I wish to acknowledge my family, and especially my parents Stewart and Carolee Smith. I can honestly say that any good I have accomplished in my life can be directly traced back to them. They are unwavering in their support, their wisdom, and their prayers. I dedicate this thesis to them, with all my love and appreciation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

On the global stage, the Portuguese language is making its way ever closer towards the forefront. According to The World Bank, Brazil (the native country of over 80% of the world's Portuguese speakers) has the sixth largest economy as of 2012 and is advancing at an impressive rate (2013; Estatística, 2013). Of all the world's nations, Brazil has the fifth largest population (Estatística, 2013). Besides Brazil, Portuguese has official status in seven countries, namely Angola, Cape Verde, Guinea-Bissau, Mozambique, Portugal, São Tomé and Príncipe, and East Timor, with significant Portuguese speaking populations in India (Goa), China (Macao), Japan, and numerous other locations across the world (Lewis, Simons, & Fennig, 2013). Portuguese is the sixth most commonly spoken of the world's languages after Mandarin Chinese, English, Spanish, Hindi, and Arabic (Crystal, 1997:289). Portuguese's increasing global presence has caused the world to take notice.

Beyond its world influence, Portuguese is of great interest to the linguistic community. "A language geographically so far-flung, spoken by over two hundred million people on four continents, could not fail to show a great deal of variation" (Azevedo, 2002:2), and it does. Historically, phonologically, lexically, and syntactically (etc.), Portuguese is incredibly rich and varied. Although classified as an SVO language (Lewis, Simons, & Fennig, 2013), Portuguese is not without a certain amount of syntatic variation among dialetcs and registers; pro-drop (Barbosa, Duarte, & Kato, 2005; Cyrino, 1993; 1994; Duarte, 1993; 1995), and word-order variations (Costa, 2000; Kato & Raposo; Silva, 2001), for example, have been investigated by numerous linguists.

Many of the previously mentioned syntactic studies have been performed by native speakers of Portuguese, meaning that they have the advantage of native intuition regarding questions of grammaticality. A non-native researcher (like myself) doesn’t have this advantage. Non-native researchers must depend upon empirical data from consultants or corpora when performing linguistic research. Even native speakers must often rely on corpora for the specific data and the linguistic standard that they provide. For this reason, corpus-based research is becoming increasingly popular in the linguistic community (Davies, 2008; McEnery & Wilson, 2001).

The principle goal of this thesis is to investigate word-order variations in Brazilian Portuguese and show how large corpora can be used in syntactic research. This is an important contribution to the literature, as a study of this type, incorporating both pragmatic analysis *and* corpus data to determine contextual motivation for variations in word order (whether or not context and pragmatic function determines word order), has not yet been performed for Brazilian Portuguese. With these goals in mind, I used The Corpus do Português (The Portuguese Corpus, hereafter referred to as the CDP) to examine some of the syntactic variations previously observed by native Portuguese-speaking linguists. The following research questions were designed both to examine Brazilian Portuguese (hereafter referred to as BP) and to test different capabilites of the corpus:

1. How frequent are null subjects in BP?
2. How frequent are covert objects in BP?
3. Which of the possible word order variations actually occur in BP, and how frequent are they?
4. How are the frequencies of 1, 2, and 3 affected by register?
5. Why do null subjects, covert objects, and word order variations occur?
6. Is *the CDP* a viable source of data for syntactic and word order research?

In order to answer these questions, samples of BP were collected from the CDP and analyzed to determine the presence and frequency of relevent syntactic phenomena: null subjects, covert objects, and variations in word order. After data was collected regarding the prevalence of these variations, comparisons were made across the different registers of the corpus (academic, newspaper, fiction, and oral BP), in order to determine if register affected the frequency of the variations.

After the existence and prevalence of word order variations was determined by the corpus data, the samples were analyzed contextually using the method that Costa (2000) used to determine whether or not pragmatic issues of *topic* and *focus* were responsible for variations. The present study was unique, in that Costa applied his analysis to European Portuguese (EP), but I examined a different dialect: BP. Also, Costa focused on theoretical contexts for his analysis, where all of the contexts analyzed in the present study were actual empirical samples collected from the corpus. The present study found that topic and focus can account for the observed word-order variations, but also that some of the variations Costa described are actually incredibly rare in BP.

The chapters that follow elaborate upon these points. Chapter 2 reviews the current literature on pro-drop and word-order variations in Portuguese, and on the use of corpora for this type of research. Chapter 3 describes the methods used in collecting and analyzing the corpus data. Chapter 4 presents the numerical results of the corpus analysis, and the statistical analyses used to determine whether or not observed variations were significant. Chapter 5 discusses the results presented in chapter 4 at great length, presenting possible explanations for the observed variations. Chapter 6 concludes this study, discussing implications, limitations, and directions for future work.

2 Review of Literature

This review of literature will provide important background information for the present study, setting it up to answer the key research questions, as stated in the introduction. First, this chapter will provide some background information about BP. Then, it will explain the idea of pro-drop, as well as studies that have examined pro-drop in BP. The third section will address word order and word order variations in both BP and EP (European Portuguese), and the final section will discuss large monitor corpora, like *the CDP*, and their use in syntactic research.

2.1 Brazilian Portuguese

Brazilian Portuguese (hereafter BP) is the variety of the Portuguese language spoken primarily in Brazil. According to the 2010 national census performed by the Instituto Brasileiro de Geografia e Estatística (the Brazilian Geographical and Statistical Institute), it is written and spoken by nearly all of the 190 million inhabitants of Brazil, as well as by several million Brazilian emigrants located primarily in the U.S., Paraguay, Japan, Portugal, and Argentina (Estatística, 2013). In his book entitled *Portuguese: A Linguistic Introduction* (2002), Azevedo describes BP and its presence in the world. According to his research, Brazilians make up 80 percent of the world's Portuguese-speaking population, meaning that even if speakers of other varieties of Portuguese weren't counted, it would still be placed seventh among the world's languages with the most speakers (Azevedo, 2002).

Brazilian Portuguese has an involved external history, which is outlined at great length in the first chapter of Azevedo's book. Suffice it to say that, since Brazil gained independence from Portugal in 1822, it has followed a very different path from other Portuguese speaking nations in Europe, Africa, and Asia. According to Azevedo, this has contributed greatly to the specific character of BP. In African countries, Portuguese is the native speech only of the minority, whereas in Brazil it has been the native language of the majority for generations. Furthermore, in Brazil, contact with languages spoken by indigenous inhabitants of South America, African slaves, immigrants and foreign cultures including French, British, American, German, Italian, and Dutch, have substantially changed BP in ways that other varieties of Portuguese have not experienced (Azevedo, 2002: 18).

Portuguese has its origins in Latin, and is considered one of the main Romance languages along with Spanish, Italian, and French, with Spanish being the most similar (both languages having originated on the Iberian Peninsula as local dialects of Vulgar Latin) (Azevedo, 2002). There are several ways in which BP is different from EP. Probably the biggest difference is in pronunciation. Like Spanish and Italian, spoken BP generally has clearly articulated vowels. EP, on the other hand, tends to weaken or eliminate untressed vowels, causing sequences of consonants that don't exist in BP. The syntactic core of the two dialects is basically the same; however, there are clear differences in certain aspects of sentence structure (Galves, 1987; 1991; Kato, 1993; 1994), including the use (or non-use in the case of BP) of unstressed pronouns. Further differences include lexical, orthographic, and pragmatic features that are inconsistent between the two varieties. This has led to much debate on the status of BP as a dialect of Portuguese, or as a separate language (Azevedo, 2002), which I will not entertain here at any length. It is important to note, however, that whether it is a dialect or a language, BP is potent, it is interesting, and it is relevant. Certainly it can be argued that BP is a worthy candidate for linguistic research.

2.2 Pro-drop and BP

In his 1981 book *Lectures on Government and Binding: The Pisa Lectures*, Noam Chomsky describes a class of languages, which he labels *pro-drop* languages. This title stems from the ability of these languages to omit certain classes of pronouns when they can be inferred pragmatically. Haspelmath (2001) describes a common type of pro-drop language in which the subject pronoun is covert (hereafter referred to as *null-subject languages* or *NSLs*), explaining that “the majority of the world's languages have bound person markers on the verb that cross-refer to the verb's subject (or agent)... In most languages [bound person markers] can occur on their own and need not co-occur with overt subject NPs” (p. 1500).

Rizzi (1982) effectively demonstrated the occurrence of the phonetically null subjects in his now famous comparison of Italian, a documented null-subject language, and English, a language which requires overt subjects (p. 117):

(1)	(a)	**Ø**	***verr-à.***
		NS	come-3SG.FUT
		‘He/she/it will come.’	
	(b)	****Ø will come.***	

As illustrated in (1a), Italian allows for a null subject. The verb *verrá* ‘will come’ has the bound, third-person singular marker -*á* which makes the intended subject of the sentence clear even in the absence of an overt subject. This shows a clear difference between Italian and English, which has no bound person markers on the verbs (*will* or *come*), meaning that a phonetically null subject in English (1b) yields an ungrammatical sentence.

Chomsky (1981) describes several properties that appear to be related to pro-drop languages, two of which are particularly exclusive to languages of the null-subject family (e.g., Italian). Example (2a) describes the ability of pro-drop languages to omit phonologically overt subjects, as illustrated previously in (1a) and again in (3a). Example (2b) describes how pro-drop languages can exhibit free inversion, meaning that the positions of the subject and the verb can be inverted with relation to each other. This is illustrated in (3b) (p. 253):

(2) (a) missing subject

(b) free inversion

He follows the tradition of illustrating these pro-drop features with Italian examples (p. 240):

(3)	(a)	***h-o***	***trovato***	***il***	***libro.***
		have-1SG.PRS	found	the	book
		‘I found the book.’			
	(b)	***h-a***	***mangiato***		***Giovanni.***
		have-3SG.PRS	eaten		Giovanni
		‘Giovanni ate.’			

In (3a), there is no overt subject in the Italian sentence, an occurrence that characteristically is not a property of non-NSLs like English or French. In (3b) the subject is inverted in the Italian sentence. Here again, this common property of NSLs is generally not a characteristic of non-

NSLs, happening only under highly restricted conditions in French. Chomsky assumes that in the case of inversion, the NP occurring in the post-verbal position is coindexed with the empty subject position, but not in a way that is relevant to binding. This means that (3b) is assumed by Chomsky to exhibit pro-drop, even though there is an overt subject. The present study will not specifically address this debate about whether or not (3b) has pro-drop, but a greater description of word order variation in pro-drop languages is in section 2.3 of this chapter.

The discussion of the pro-drop languages exhibiting the null-subject property is relevant to BP, as BP is a null-subject language, meeting the criteria established by both Chomsky (1981) and Rizzi (1982) as described in (2). This is illustrated in (4) using samples from the fiction portion of the CDP with translations from Larousse (2008):

(4)	(a)	***atravess-amos***	***um***	***período***	***estacionário.***
		go.through-1PL	a	period	stationary
		‘We went through a stationary period.’			
	(b)	***esbat-iam-se***	***as***	***nuvens.***	
		faint-3PL.PST-REFL	the	clouds	
		‘The clouds fainted.’			

In (4a), the BP sentence has no overt subject where the English translation requires the first person plural pronoun *we*. This is exactly parallel to the Italian examples of omitted subjects provided by Rizzi and Chomsky in (1a) and (3a) respectively. The BP example in (4b) shows subject inversion in the same way that inversion occurred in Chomsky’s Italian example in (3b). For these reasons, the present study will categorize BP as a null-subject language, although it is the researcher’s belief that BP might be more accurately classified as a quasi-null-subject language. It appears to have a much stronger null-subject tendency than English, but certainly less than other languages like Italian or Spanish. Unfortunately formalist linguistics typically classifies languages as belonging to one category or the other (Chomsky, 1981; Rizzi, 1982; Haspelmath, 2001)

The literature has also categorized Portuguese as a null-subject language. In Duarte's *O sujeito pronominal no português coloquial europeu* (The pronominal subject in colloquial European Portuguese), she performed a study to describe the null-subject properties of European Portuguese. A small corpus was created using samples collected from interviews conducted with 30 speakers of EP from two different age groups. She looked at samples where the subject was anaphorically related to the subject of a preceding sentence, and samples where there was no relation. The results of her study are included in Table 2-1, (Duarte, 1995:8):

Table 2-1: Null Subjects in EP

Person	*Anaphor Subject* *Null / Total / (%)*	*Non-Anaphor Subject* *Null / Total / (%)*
1st	334 / 561 / (60)	243 / 459 / (53)
2nd	101 / 138 / (73)	96 / 133 / (72)
3rd	303 / 417 (73)	194 / 305 / (64)

In Table 2-1, Duarte's data shows that in colloquial EP, the majority of speech samples have null subjects. Duarte concluded that for all persons, in sentences with and without anaphoric relation, spoken EP always prefers the null subject.

The literature also specifically classifies BP as a traditionally null-subject language (Barbosa, Duarte, & Kato, 2005; Duarte, 1993; 1995; Galves, 1991). Regarding the prevalence of the null-subject property in Brazilian Portuguese, several studies of note have been performed.

Duarte (1993) found that spoken BP is gradually increasing in its use of overt pronominal subjects (supporting the argument for a *quasi-null-subject* classification). According to the study, EP (as stated previously) traditionally uses a null subject when it is anaphorically related to the matrix subject in a biclausal structure as illustrated in (5a) (Duarte, 1995:10), but spoken BP appears to be leaving this trend, as shown in (5b) and (5c) (Barbosa, Duarte, & and Kato, 2005):

(5) a. ***ele$_i$ não ganh-a mal, mas para aquilo que a gente***
he$_i$ NEG earn-3SG.PRS bad but for that which the people
quer Ø$_i$ ganh-a pouco.
want.3SG.PRS NS earn.3SG.PRS little
'He$_i$ doesn't do poorly, but for the things we want, (he$_i$) earns very little.'

b. ***e ele$_i$ precis-ou ir ao banheiro. Quando ele$_i$ viu***
and he need-3SG.PST go to.the bathroom. when he$_i$ saw
o que que era o banheiro, ele$_i$ fic-ou apavorado
the what that be.3SG.PST the bathroom he$_i$ be-3SG.PST terrified
'And he$_i$ had to go to the bathroom. When he$_i$ saw what
the bathroom looked like he$_i$ was terrified.'

b. ***[a casa]$_i$ vir-ou um filme quando ela$_i$ teve de ir abaixo***
the house$_i$ turn-3SG.PST a movie when it$_i$ had of go down
'The house$_i$ became a movie when it$_i$ was demolished.'

As shown in the EP example in (5a), the subject in the final clause was overt, because it was anaphorically related to the subject in the preceding clause. In the BP samples in (5b) and (5c), the subject in the same context is not omitted. In (5b), the subject of the final clause *ele* 'he' would have been omitted in EP, as it is anaphorically related to both the subject of the preceding clause and the subject of the preceding sentence. In the BP sample in (5c), the subject of the final clause *ela* ‘she/it’ would have been omitted in EP, since it is anaphorically related to the subject *a casa* ‘the house’ of the preceding clause.

In Duarte’s 1993 study, the data showed a significant increase in overt pronominal subjects in BP over the past century. She examined text from seven popular plays, one from each of the periods represented in Figure 2-1, which was adapted from this study and illustrates the increase in overt subjects (p. 112):

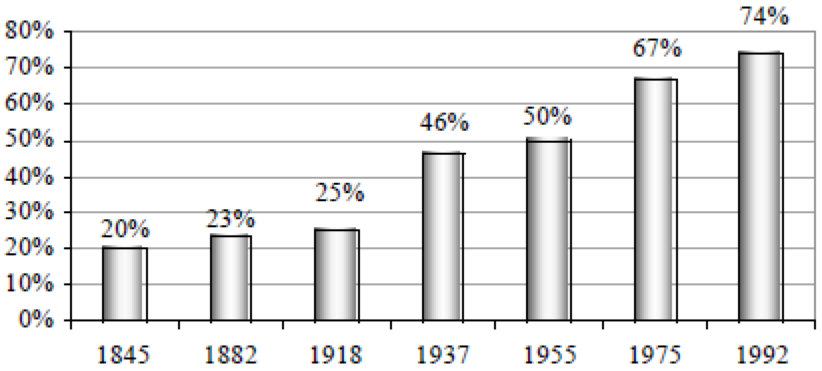


Figure 2-1 Overt pronominal subjects through seven periods (Duarte, 1993: 112)

It was unclear from her article exactly which sentences she examined from the plays, but for the samples examined, Duarte’s findings in Figure 2-1 show that the rate of overt pronominal subjects in the first half of the 1800s was 20%, but by the end of the 1900s it had increased to 74%.

While this is a compelling diachronic study of null-subject behavior in the oral register of BP, there are some possible weaknesses in the data examined. First, the corpus analyzed came from popular plays written in the different periods examined. While examining written language lends some credence to the idea that there is a change in pro-drop, as writing samples reflect the writers' perception of the language, it is difficult to accept that written language, even in a form that attempts to mimic the oral register such as a play, is a legitimate representation of actual oral language. Duarte confirmed the data of the most recent play (1992) by examining oral data produced by college-educated and middle-level educated adult speakers in Rio de Janeiro. A corpus was created using recorded interviews of the 13 consultants. It was found that 71% of the samples analyzed had an overt subject (Barbosa, Duarte, & and Kato, 2005). While confirming the play data with data produced by real speakers lends credibility to the findings of the previous study, it still seems to be a very narrow scope for data collection. The only data that was truly oral was collected from *one* demographic in a very specific region, meaning that it may not represent the general behavior of the language. Perhaps a broader study with a larger quantity of more diverse data that came from truly oral sources would provide more convincing evidence of any changes in null-subject behavior in BP.

In 2005, Barbosa, Duarte, and Kato sought to confirm whether or not written BP is losing the null subject in the same way that spoken BP appears to be. They examined a written corpus containing magazine interviews that were enclosed in Sunday editions of newspapers from Lisbon (*O Púbico*) and Rio de Janeiro (*R*evista *Domingo*) in 1999 and 2000. With this corpus, they compared both EP and BP. Their results are seen in Table 2-2 (p. 13):

Table 2-2: Occurrences of null and overt subjects in EP and PB (Barbosa, Duarte, & and Kato, 2005: 13)

Variety	*Null subjects*	*Overt subjects*	*Total*
EP	126 (78%)	36 (22%)	162 (100%)
BP	63 (44%)	79 (56%)	142 (100%)

As shown in Table 2-1, they found a significant difference in the prevalence of null subjects in EP and in BP, which they had expected.

Here again, these results are very interesting, but there are some points where their data might not be showing the whole picture. Notice that they had only one source of data for each variety of Portuguese. This is a very narrow look at the language, making it difficult to make any sort of general descriptive claims as to the null-subject properties of Portuguese. Also, they had less than 200 samples for each variety. Perhaps this study could be improved with a much larger and varied data set, since one magazine is hardly sufficient to meaningfully represent an entire dialect.

While Portuguese is a confirmed null-subject language, it has an additional pro-drop characteristic that is rarer and less discussed in the literature. It has been observed that in BP, objects can also be omitted (Cyrino, 1993; 1994), as shown by the examples in (6a) and (7a) taken from the newspaper register of the CDP. (6b) and (7b) show the examples as they would be if the objects were expressed overtly. The translations from (6) and (7) come from Larousse (2008). This feature will be referred to as the *covert object* and will be abbreviated as CO:

(6) a. ***o consumo hoje assusta Ø.***
the consumption today scares CO
'Today's consumption rates frighten' (lit.).

b. ***o consumo hoje assusta eles.***
the consumption today scares them
'Today's consumption rates frighten them.'

(7) a. ***naturalmente a gente não procurava Ø.***
naturally the people no sought CO
'Naturally we didn't seek' (lit.).

b. ***naturalmente a gente não procurava ele.***
naturally the people no sought him
'Naturally we didn't seek him.'

Probably the most important of the few studies that have been done regarding covert objects in BP was performed by Cyrino (1993; 1994). She did a diachronic study in which she collected data from BP texts representing five centuries. Her goal was to examine different types of covert objects in the oral register of BP and how they have changed over time. Just like Duarte (1993) she examined mostly plays, as they attempt to approximate natural speech. Here again, while plays may imitate speech, they don't actually constitute naturalistic data. Unlike Duarte, she did not confirm her results by examining recorded speech samples, as she was more concerned with the diachronic changes in covert objects over the centuries. She gathered 300 samples from each of the last five centuries and examined which types of covert objects occurred. The frequency results that she gathered from her play corpus are shown here in Table 2-3 (Cyrino, 1994:169):

Table 2-3: Occurrences of covert objects in BP over five centuries

	Covert Object		*Overt Object*		*Total*	
Century	*#*	*%*	*#*	*%*	*#*	*%*
1500s	31	10.7	259	89.3	290	100
1600s	37	12.6	256	87.4	293	100
1700s	53	18.5	234	81.5	287	100
1800s	122	45.0	149	55.0	271	100
1900s	193	79.1	51	20.9	244	100

Cyrino not only found that covert objects exist in BP, but that they are increasing in frequency (as shown in Table 2-3). While this is compelling, there are some weaknesses to her study which need to be addressed in future research. First, her "oral" data is all gathered from plays, which, as previously mentioned, is not truly naturalistic oral data. Second, her corpus is limited both in the number of registers (only plays and similar texts) and in size (around 300 samples per century). It is clear from the corpus and from Cyrino's data that there are instances where the object is dropped in BP, but due to the limited nature of the principle studies (as with the null-subject feature) much work still needs to be done to determine the true frequency for these pro-drop constructions in the most relevant registers of BP.

2.3 Word Order and BP

Word order is an important topic in linguistics, being one of the key ways that languages are classified. The term *word order* usually refers to the location of the subject, object, and verb of the sentence in relation to each other. With regards to these three key linguistic components, there are six logically possible word orders for languages (SVO, SOV, VSO, VOS, OSV, and OVS), with the additional possibility of a free word order (Costa, 2000). The literature talks at great length about general word-order variations (Biber et al., 2002; Downing & Noonan, 1995; Longman, 1999; Rooth, 1985; Trujillo)

In Silva's dissertation, later published as the book *Word Order in Brazilian Portuguese,* she describes the word order of BP. First, she explains that most declaratives in BP exhibit the SVO word order, stating that this is the default order for sentences involving transitive verbs. She provided the following examples (p. 2):

(8) a. ***a Ana compr-ou muita coisa nesta loja.***
The Ana buy-3SG.PST much thing in.this store
'Ana bought much stuff in this store' (lit.).

b. ****Compr-ou a Ana muita coisa nesta loja.***
buy-3SG.PST the Ana much stuff in.this store
'Bought Ana much stuff in this store' (lit.).

c. ****Compr-ou muita coisa a Ana.***
buy-3SG.PST much stuff the Ana
'Bought much stuff Ana' (lit.).

Silva describes (8b) and (8c) as ungrammatical, reiterating that only the SVO option shown in (8a) is acceptable for transitive verbs in BP. For sentences containing unaccusative verbs, she states that it is possible for there to be a verb-subject order. She provides the examples shown in (9) (p. 3):

(9) a. ***a Maria cheg-ou.***
the Maria arrive-3SG.PST
'Maria arrived.'

b. ***cheg-ou a Maria.***
arrive.3SG.PST the Maria
‘Arrived Maria’ (lit.).

In (9), both SV and VS are described as grammatical options. It is not surprising that subject-verb inversion should occur in BP, as a common property of null-subject languages is subject-verb inversion (Chomsky, 1981; Rizzi, 1982; Barbosa, Duarte, & Kato, 2005) (see section 2 of this chapter).

Silva goes on to explain that BP exhibits some interesting features that make it different for other Romance languages, even European Portuguese. For example, BP does not allow for a postverbal subject in interrogatives, where other Romance languages require subject-verb inversion in questions. Silva states that in BP, this inversion yields ungrammatical sentences, and that SVO is the order used in both interrogatives and declaratives. The following examples in (10) are adapted from Silva (2001) and demonstrate this phenomenon (p. 4):

(10) a. ***o que o Paulo compr-ou?***
the what the Paulo buy-3SG.PST
‘What did Paulo buy?’

b. ****o que compr-ou o Paulo?***
the what buy-3SG.PST the Paulo
‘What bought Paulo?’ (lit.)

Kato and Raposo also described some interesting features of word order in their paper *European and Brazilian Portuguese Word Order: Questions, Focus and Topic Constructions.* In this paper, they contrast word order variations in BP with those in European Portuguese. The examples in (11) are particularly interesting. Kato and Raposo indicate that the motivation for these other variations is focus and topic (p. 267-268):

(11) a. ***quem com-eu o bolo?*** (EP/BP)
who eat-3SG.PST the cake
'Who ate the cake?'

b. O:Topic V S:Focus
(o bolo) com-eu A MARIA. (EP/*BP)
the cake eat-3SG.PST the Maria
'(The cake) ate Maria' (lit.).

c. O:Topic S:Focus V
(o bolo) A MARIA com-eu. (*EP/BP)
the cake the Maria eat-3SG.PST
'(the cake) Maria ate' (lit.).

(12) a. ***Quanto cust-ou o seu carro?*** (EP/BP)
how.much cost-3SG.PST the your car
'how much did your car cost?'

b. S:Topic V O:Focus
(o carro) cust-ou-me $5,000. (EP/*BP)
the car cost-3SG.PST-me $5000
'(The car) cost me $5000.'

c. O:Focus V S:Topic
$5,000 me cust-ou o carro. (*EP/BP)
$5000 me cost-3SG.PST the car
'$5000 me cost the car' (lit.).

(13) a. ***a Maria recomend-ou-me ESTES DISCOS*** (EP/*BP)
the Maria recommend-3SG-me these records
'Maria recommended me these discs' (lit.).

b. ***ESTES DISCOS a Maria me recomend-ou.*** (*EP/BP)
these records the Maria me recommend-3SG-PST
'These records Maria me recommended' (lit.).

According to the authors, these examples demonstrate that in both BP and EP it is possible to have a fronted topic in a left-dislocated position ((11b, c), (12b)). The big difference demonstrated between the dialects is that in BP, a definite NP may be a marked focus in pre-verbal position ((11c), (12c), and (13b)). Rooth (1985) defines marked focus as when a lexical item receives prosodic emphasis in an utterance (intonation). According to Rooth, this prosodic marking invokes a set of possible alternatives from which a particular one is specified. In EP a definite focused NP must be in the unmarked post-verbal position, meaning that the unmarked position for focus is post-verbal ((11b), (12b), and (13a)) (Kato & Raposo, 1996). Their samples show situations where other word orders besides SVO may occur (OSV in (11c) and (13b), and OVS in (12c). They showed several word orders that Silva didn't address, but it is important to note that Silva was working with a neutral context. It is interesting that Kato and Raposo did not provide any explanation as to their method of data collection. They state that their examples are grammatical, using them as evidence for certain variations, but some of their examples seem strange, particularly (12c) and (13b).

Costa (2000) also did a great deal to explain Portuguese word-order variations in his work *Word Order and Discourse-Configurationality in European Portuguese*. He focused principally on European Portuguese as the title states, but was much more open regarding the grammaticality of the different variations. He provided the examples in (14) (p. 94):

(14) a. ***o Paulo com-eu a sopa***
the Paulo eat-3SG.PST the soup
'Paulo ate the soup.'

b. *Comeu o Paulo a sopa*

c. *Comeu a sopa o Paulo*

d. *A sopa comeu o Paulo*

e. *A sopa o Paulo comeu*

f. **O Paulo a sopa comeu*

According to Costa, only (14f), exhibiting the SOV word order, is ungrammatical. Like Kato and Raposo, Costa suggested that variation is not discourse-neutral, but the reflex of discourse-configurationality, meaning that topic and focus are important for determining the felicitous word orders for a given context. In order to determine topic and focus, he used the following tests (p. 103-104):

1. In a questions-answer pair, a focused constituent in the answer replaces the *wh*-word in the question.
2. A topic is information already referred to in the discourse or a subpart of a given referent.

It is important to note that there are actually two types of focus: identification focus and new information focus (Kiss, 1998). Costa did not specifically make the distinctions, but he seems to be referring to new information focus in his paper. The present study, like Costa, is more interested in new information focus. Costa examined each word order to determine possible contexts. His word order tests are valuable to the present study, as they are a model for determining if word-order variations that occur in BP are dependent upon the same pragmatic contexts.

2.3.1 Context for SVO (the prevalent word order)

According to Costa, the SVO order with definite subjects may be used in either of two cases: the subject is familiar to the discourse participants but the object is not, as seen in (15), or both subject *and* object are familiar (16) (p. 104):

(15) (a and b are checking which languages each person in a given group speaks. They are talking about Paulo.)

a. ***o Paulo sab-e que línguas?***
the Paulo know-3SG.PRS what languages
'Paulo knows which languages?'

b. ***o Paulo sab-e francês.***
the Paulo know-3SG.PRS French
'Paulo knows French.'
****Sabe o Paulo francês.***
****Sabe francês o Paulo.***
****Francês o Paulo sabe.***
****Francês sabe o Paulo.***

According to Costa, only SVO was legitimate for EP within the context. None of the other orders were felicitous.

(16) (a and b are checking which persons in a given group speak French. They are talking about Paulo.)

a. ***o Paulo sab-e francês?***
the Paulo know-3SG.PRS French
'Paulo knows French?' (lit.)

b. ***o Paulo sab-e francês.***
the Paulo know-3SG.PRS French
'Paulo knows French.'
****Sabe o Paulo francês.***
****Sabe francês o Paulo.***
Francês o Paulo sabe.
****Francês sabe o Paulo.***

Here again, SVO is legitimate. The only difference between (16) and (15) is that in (16), the OSV word order is also possible if the object (French) is topicalized. Costa points out that just because the subject is old information does not mean it is the topic (Buhring, 1995), which he illustrates by the fact that it is not in complementary distribution with a topicalized constituent in (17) (p. 104-105):

(17) a. ***com quem é que o Paulo falou sobre***
with what be.3SG.PRS that the Paulo speak-3SG.PSTabout
o Big Bang?
the Big Bang
'With whom Paulo talked about the Big Bang?' (lit.)

b. ***sobre o Big Bang, o Paulo falou com o Pedro.***
about the Big Bang the Paulo speak.3SG.PST with the Peter
'About the Big Bang, Paulo talked with Pedro' (lit.).

In (17B), *o Big Bang* 'the Big Bang' is the topicalized constituent. Both 'the Big Bang' (the indirect object), and Paulo (the subject) constitute old information, but in (17B) the indirect object has been placed first, in the topic position. As shown in the example, it is possible for more than one constituent to consist of old information, but there will only be one topic. The topic will be the one that occurs in the first position (Costa, 2000). Costa explained that SVO order is also acceptable with indefinite subjects if they are not new information, as demonstrated in (18) (p. 105):

(18) a. ***est.ão imensos animais neste parque: cães gatos galinhas.***
be-3PL.PRS immense animals in.this park dogs cats chickens
'There are a lot of animals in the park: dogs, cats, chickens.'

b. ***olha: um cão mord-eu uma criança.***
look a dog bite-3SG.PST a child
'Look: a dog bit a child.'
****Mordeu um cão uma criança.***
****Mordeu uma criança um cão.***
****??Uma criança um cão mordeu.***
****Uma criança, mordeu um cão.***

If the indefinite subject represents new information, the SVO order is not felicitous (p. 105):

(19) a. ***o que é que mord-eu o Paulo?***
the what be.3SG.PRS that bite-3SG.PST the Paulo
'What bit Paulo?'

b. ****uma cobra mord-eu o Paulo.***
a snake bite-3SG.PST the Paulo
'A snake bit Paulo.'

Costa concluded from the preceding examples that preverbal subjects must constitute old or accessible information (Costa, 2000).

2.3.2 Context for other word orders

Costa described contexts for which other word orders are felicitous in European Portuguese. He explained that the VSO order is best when both subject and object are new in the discourse, as demonstrated in (20) (p. 105):

(20) a. ***ninguém sab-e línguas neste grupo.***
no.one know-3SG.PRS languages in.this group
'No one in this group knows any language.'

b. ***sab-e o Paulo francês.***
know-3SG.PRS the Paulo French
'Knows Paulo French' (lit.).
****O Paulo sabe francês.***
****Francês o Paulo sabe.***
****Sabe francês o Paulo.***
****Francês sabe o Paulo.***

He then discusses utterances in which only the subject is new information. In this case (21) the only felicitous orders are VOS or OVS, derived by object left-dislocation, which, Costa states, is not surprising, since the object is old information (p. 106):

(21) a. ***ninguém sab-e francês neste grupo?***
no.one know-3SG.PRS French in.this group
'No one knows French in this group?' (lit.)

b. ****sab-e o Paulo francês.***
know-3SG.PRS the Paulo French
'Knows Paulo French' (lit.).
****O Paulo sabe francês.***
****Francês o Paulo sabe.***
Sabe francês o Paulo (não sabe?).
Francês sabe o Paulo.

Costa concludes from the examples in this section that:

1. Preverbal definite subjects are old information.
2. Preverbal indefinite subjects are old information.
3. Postverbal subjects must be new information:
    a. If they precede the object, the object is also new information.
    b. If the object is not new information, the subject follows it (Costa, 2000: 106).

His overall conclusion is that word-order variation in European Portuguese is not free. He states that each order reflects a different discourse function. These findings agree with previous research, which proposes that subject position is related to discourse function (Ambar, 1992; Martins, 1994; Costa, 1996a; 1996b). The literature also supports the proposition that object position is determined by discourse function (Cinque, 1992; Nash, 1995; Reinhart, 1995). For both subject and object position, the literature suggests that new information generally receives prosodic focus (Jackendoff, 1972; Rooth, 1985), and according to Cinque's theory of sentence stress assignment, the most embedded constituent, typically the last constituent of the utterance, receives the nuclear stress (Cinque, 1992; Costa, 1996c). This means that whichever constituent is the focus (new information) will occur in the most embedded position at the end of the utterance, unless some other position is marked prosodically as described by Rooth (1985), allowing the focus to occur in a different position (Kato and Raposo, 1996).

The studies performed by Kato and Raposo, and by Costa both agree with the literature, suggesting that topic and focus play an important role in determining Portuguese word order. A comparison of their findings for BP and EP is found in Table 2-4 (Costa, 2000; Kato & Raposo, 1996):

Table 2-4: BP and EP word order by topic and focus

Word Orders	*Kato and Raposo (1996)*		*Costa (2000)*
	BP	*EP*	*EP*
SVO	-	S:Topic V O:Focus	S:Topic V O:Topic/Focus
SOV	-	-	*S O V
VSO	-	-	V S:Focus O:Focus
VOS	-	-	V O:Topic S:Focus
OSV	O:Topic S:**FOCUS** V	-	O:Topic S V
OVS	O:**FOCUS** V S:Topic	O:Topic V S:Focus	O:Topic V S:Focus

The two studies are in agreement about the effects of topic and focus on word order in EP, although Costa (2000) presented far more possibilities than Kato and Raposo did (1996). The shaded cell indicates a point where BP and EP are different according to the literature. As shown in Table 2-3, in EP, the focused constituent is always in the most embedded position, except when specially marked prosodically (specially marked focus is labeled with bold, upper case type in Table 2-4, as shown in the OSV and OVS sections of BP). Only two word orders were presented for BP. The OSV order in BP appears to be similar to that of EP presented by Costa, in that both have a topicalized object. Kato and Raposo specified that in BP the subject can be focus in OSV sentences. The biggest difference between the two varieties of Portuguese was with the OVS sentences. In BP, Kato and Raposo determined that the preverbal object is focus, and the postverbal subject is topic. This was illustrated in (12c) and (13b) of this chapter. It is important to note that the preverbal objects in these examples are prosodically marked as the focus of the sentence (Kato & Raposo, 1996; Rooth, 1985).

2.4 Corpora and Syntax Research

The previous sections of this chapter are important in that they describe several interesting features of BP (pro-drop, word order, etc.) and how they have been researched in the past. This has set the foundation for many of the research questions of the present study. What remains is to discuss different linguistic methodologies and determine which is best for answering the present questions regarding the frequency of syntactic variations in pro-drop and word order. There are two principle types of methodologies: empirical and formal. The debate between proponents of empirical methods (fieldwork, corpus linguistics, etc.) and the formalist method has been a long one. The aim of the present study is to show that using large corpora is an effective means of syntactic research. For this reason, this section will describe the debate between the two parties and illustrate how modern corpus methodologies can solve problems associated with the oldest forms of empirical data gathering, answering objections raised by Chomsky (1957), and providing a more grounded, data-driven option to intuition-based research.

Regarding the two traditional methodologies, in his 1957 work *Syntactic Structures*, Chomsky described what would become formalist linguistics (p. 12):

> "The fundamental aim in the linguistic analysis of a language L is to separate the *grammatical* sequences which are the sentences of L from the *ungrammatical* sequences which are not sentences of L, and to study the structure of the grammatical sequences...For the purposes of this discussion, however, suppose that we assume intuitive knowledge of the grammatical sentences...and ask what sort of grammar will be able to do the job of producing these in some effective and illuminating way."

Chomsky's idea was to use intuition, rather than large quantities of empirical data, to determine which possible utterances were grammatical, and therefore how a particular "grammar" of a language behaved. This formalist method, based on intuition, is commonly used today, as illustrated in the previous sections of this chapter.

The alternative to the formalist approach is empirical data collection (McEnery & Wilson, 2001). Prior to Chomsky's *Syntactic Structures* (1957), most linguistic research was very data-driven. It even used corpora. Early linguists worked with Native American languages for example, and, having no native speaker intuition, the only way to work with these languages was to obtain and carefully organize as much data as they could. Empirical linguistic research is strongly influenced by *positivism* and *behaviourism* which includes the idea that if linguistics could collect enough data to model the input a language learner would receive (L1 or L2), they would be able to predict the language development of that speaker; therefore, the collection of large amounts of language samples is crucial (Davies, 2008; Bednarek & McCarthy, 2011). Modern corpus linguists ascribe to this same theory, using large corpora to model the target language and its behaviors.

Since the Chomskyan revolution, there has been a strong debate between proponents of the two methodologies, as “Chomsky forcefully attacked many of the methodological underpinnings of previous corpus-based and empirically-based research” (Davies, 2008: 150). One of the priniciple points of criticism was that the databases that linguistis created in the past were much too small to be useful. Chomsky demonstrated that even a million word corpus would provide data for some linguistic phenomena that was much too sparse to actually provide insight into actual language processing. Chomsky argued that much of the corpus data of the time was trivial, providing random factual information about the world, but little about language itself. Possibly his biggest argument was in favor of introspection, rather than the examination of a massive database. He thought that it made more sense to sit down with a native speaker, or to probe one’s own intuitions if one were a native speaker in order to more quickly and easily obtain relevant data. According to Davies (2008:150), Chomsky’s critiques of data-based linguistics sent corpus linguistics underground for the next 20-30 years.

It wasn’t until the 1980s when the true interest in corpus linguistics was rekindled. Many researchers in the 1980s began calling for a more nuanced model of grammar that broke away from binary judgments on grammaticality, and adopted more of a tendency approach, which works incredibly well with the corpus approach (Davies, 2008: 151). Today, both methodologies exist, but even with the availability of large corpora, much syntactic research still follows the formalist tradition.

For the present study, it is important to understand how modern corpus linguistics both corrects the problems associated with older empirical methods and provides a valid alternative to the formalist method. Davies (2008) concedes that Chomsky may have been correct in stating that small corpora (one million words, etc.) were of little value; however, with the 1980s came advancements in technology that made it possible to create much larger corpora (hundreds of millions of words), making Chomsky's earlier objection irrelevant. Researchers further showed that Chomsky's second objection regarding the triviality of corpus data is not true at all, especially for the programming of computers for natural language processing (Davies, 2008; McEnery & Wilson, 2001).

Davies addresses a fourth factor of particular importance in the debate between formalists and corpus linguistics: formal linguistics tends to emphasize the natural primacy of linguistic intuition. This is a problem, according to Davies, because researchers would ignore empirical data that showed their theories to be flawed. Researchers would argue that in *their* dialect (or even idiolect) the data was exacly how they claimed. Therefore, a standard that can check introspective data is necessary, and large, publicly-available databases are the solution (Davies, 2008: 152).

This debate between formalist and corpus methodologies is key in the present study, as the problems addressed by both sides of the debate are visible within the syntactic research presented in this review. Chomksy argued that small data samples provide sparse information of little value. Regarding the null-subject and covert-object research presented in section 2, much research has been done (Barbosa, Duarte, & Kato, 2005; Cyrino, 1993; 1994; Duarte, 1993;

1995), but in all of these studies, either the source of the data (plays in Duarte's dissertation (1993) and Cyrino's research (1993; 1994), or two magazines in Barbosa, Duarte, & Kato, (2005)) or the quantity of samples (less than 200 in Barbosa, Duarte, & Kato, (2005) and 300 per century in Cyrino (1994)) was severely limited in such a way as to prevent any claims as to the general trend of the language, just as Chomsky claimed. Furthermore, all of these studies examined only one register of BP, which is clearly not representative of the language as a whole. The solution to limited database size is a large, publicly available corpus with millions of samples from various sources.

Silva (2001), Kato and Raposo (1996), and Costa (2000), are examples of the other extreme in syntactic research: the formalist approach. They take the grammaticality (or ungrammaticality) of word-order variations for granted, by providing hypothetical samples and contexts, then determining the grammaticality of these samples by their own intuition as native speakers (a common practice in theoretical syntax). No empirical data was presented in these studies to show these phenomena. While they are native speakers, and naturally have good intuitions about grammaticality in Portuguese, it is necessary to see which word orders actually occur in the language, outside of the theoretical. Relying solely on intuition can lead to problems in research. First, native speakers can make errors when asked about grammaticality, or they can be inconsistent in their judgments (Nagata, 1988; Takaie, 2002). Second, linguists traditionally ask the question "is this grammatical?" and that is their principal focus. They don't always think about context or discourse function (as in Silva's 2001 study in which context was never discussed and assumed to be neutral); therefore, their data is not naturalistic. Here again, a large corpus consisting of millions of samples of real, natural data is the solution.

The word order research in section 3 is a very clear example of these problems, and could be greatly improved by empirical corpus data. The studies didn't even mention a method of data collection. The syntacticians merely provided their own examples with their own intuition as to the grammaticality of the variations. While their analysis is interesting, and relevant to the topic of the present study, there is a definite disconnect between their theoretical discussion and real-world BP. More theoretical papers avoid the complications of data collection and focus entirely on analysis of their own examples. This calls into questions the legitimacy of their claims regarding the actual, natural behavior of the language. This was the main point in support of corpus linguistics. A large, public corpus serves as a standard, removing any problems associated with personal intuition (Davies, 2008). This is even more relevant for the present study as I am not a native speaker of BP, and therefore could not claim native intuition anyway.

Fortunately, a corpus has been created that solves some of the problems inherent in traditional syntactic research in BP: *the CDP.* Davies (2008) describes the characterstics of a good corpus that were taken into account when creating this corpus (p. 162):

1. Size: useful corpora typically contain tens of millions of words of text
2. Representativity: the best corpora will contain texts from a wide range of genres
3. Annotation: the texts will be lemmatized and will be tagged for part of speech
4. Architecture and interface: it will be possible to search by substring (for morphology), lemma and part of speech (for syntax), collocates and synonyms (for semantics), and frequency in different historical periods and registers (for lexical research and for stylistics and historical linguistics)

The CDP meets all of these criteria, in that it contains 45 million words (20 million words from the 1900s, 10 million from the 1800s, and 15 million from earlier centuries). For the 20 million words from the 1900s, it has 2 million words from spoken, 6 million from fiction, 6 million from newspapers and magazines, and 6 million from academic. In addition, it is divided evenly between texts from Portugal and Brazil, both overall and for each of the four registers just mentioned. This corpus is fully annotated (lemmatized and tagged for part of speech). Its architecture allows for a broad spectrum of queries, including word, phrase, substring, part of speech, lemma, synonyms, customized lists, word comparisons, collocates, and frequency-based queries (Davies, 2008).

For the reasons described above, the CDP has the potential to provide valuable information that will answer the primary questions of the present study: it can be used to determine the prevalence of null subjects and covert objects in BP. It can also be used to investigate word order variation by producing a large and varied number of samples of BP. Finally, the fact that it is organized by register makes it possible to compare these variations across registers. Should the CDP prove useful in answering these questions, it will add to the argument that large, online corpora of this type can be useful for syntax research, overcoming the downfalls in intuition research, and the smaller data samples of other empirical studies.

One criticism that has been raised about these large, online corpora is related to the relatively small amount of textual context available for each sample. These corpora include many texts that do not come from public domain sources, making them problematic when it comes to copyright and access. Without millions of dollars to obtain copyright permission from all of the text sources, the full-text version of these corpora can never be legally released into the public domain. It is possible, however, to show the node word(s) surrounded by 40-60 words of textual context (180-200 words in expanded view) in accordance with U.S. copyright laws. This is a small enough percentage of the text that the user cannot re-create the original text by putting the different pieces of the text together. Legally this is the best alternative (Davies, 2010). The present study will determine if the legal limitations placed upon the textual context availability of the corpus samples poses a problem for this type of syntax research.

3 Methodology

This chapter will explain the methods used to gather data on word order variations in Brazilian Portuguese. Sample data were gathered from *the CDP* in an attempt to answer the primary questions of this study.

3.1 The corpus

The CDP is a 45 million word corpus which contains almost 57,000 Portuguese texts from the 1300s to the 1900s. There are 20 million words from the 1900s. There are four registers included in the corpus: academic (6 million words from the 1900s), newspaper (6 million from the 1900s), fiction (6 million words from the 1900s), and oral (2 million words from the 1900s). The academic register consists of academic journals and textbooks. The newspaper register is sampled from a variety of newspapers. The fiction section of the corpus was drawn from literature from the relevant time period, and the oral register was taken from transcripts of unscripted speech from radio and television interviews, as well as many one-on-one conversations done for the purpose of creating the oral corpora of the CDP. For the 1900s, each of the registers drew equally from Brazilian Portuguese and European Portuguese samples (Davies & Ferreira, 2006).

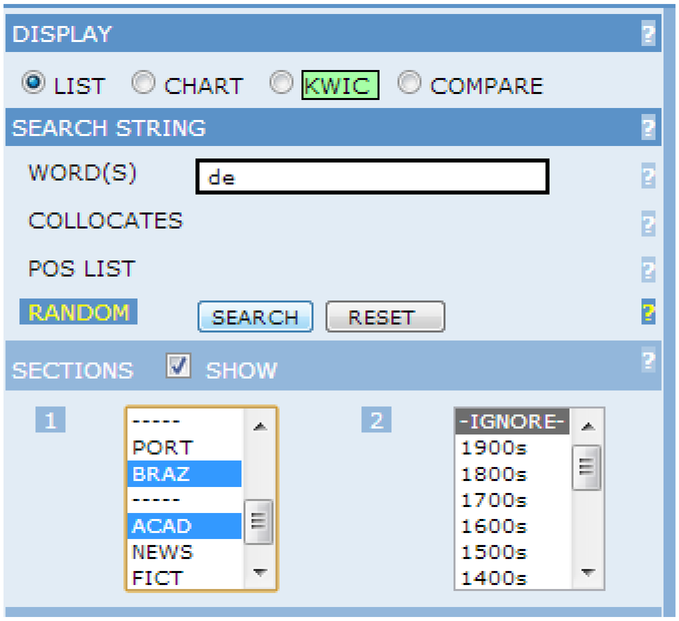
The corpus interface allows researchers to search for exact words or phrases, lemmas, part of speech, or even collocates within a ten-word window (e.g. all nouns somewhere near *cadeia* 'jail'*,* all adjectives near *mulher* 'woman', or all nouns near *girar* 'rotate'). The corpus also allows for the easy comparison of the frequency of and distribution of words, phrases, and grammatical constructions across texts, in at least three ways:

1. By register: comparisons between oral, fiction, newspaper, and academic Portuguese
2. By dialect: comparison of European and Brazilian Portuguese
3. By historical period: compare different centuries from the 1300s to the 1900s

This corpus is also useful in that it allows the researcher access to several lines of extended context before and after each token retrieved by a given search .

3.2 Procedure

To effectively determine the prevalence of covert subjects, objects, and different word orders in Brazilian Portuguese, it was necessary to perform a search that would provide a random cross-section of the language. The CDP is not designed to retrieve specific types of sentences (it isn't tagged for ditransitive sentences or wh-questions for examples); therefore, in order to produce a random list of transitive sentences, the preposition *de* (of/from) was used as the search item, as it is one of the most commonly occurring and least syntactically limiting words within the language. The word *de* is not one that is necessarily found in transitive sentences; it merely serves to limit the amount of data returned by the corpus search. The exact search method and parameters used in the search are shown in Figure 3-1.



DISPLAY

● LIST ○ CHART ○ KWIC ○ COMPARE

SEARCH STRING

WORD(S) de

COLLOCATES

POS LIST

RANDOM SEARCH RESET

SECTIONS ☑ SHOW

1: -----, PORT, BRAZ, -----, ACAD, NEWS, FICT

2: -IGNORE-, 1900s, 1800s, 1700s, 1600s, 1500s, 1400s

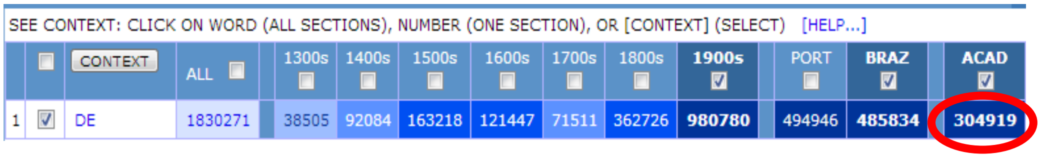
Figure 3-1 Setting the search filters in the *Corpus do Português*

Figure 3-1 shows the search functions available in the CDP. On the corpus site, the left side of the page allows the researcher to determine exactly the type of search to be performed. Here, the program was set to display a list of instances of the word *de*. Figure 3-1 further illustrates the option to filter the search so that it only shows results from specific sections of the corpus. In this case the filters were set to retrieve only samples containing the word *de* from the Brazilian dialect (BRAZ) in the academic register (ACAD). In this study, similar search filters were used to perform four different searches:

1. Samples including *de* from the academic register of BP in the 1900s
2. Samples including *de* from the news register of BP in the 1900s
3. Samples including *de* from the fiction register of BP in the 1900s
4. Samples including *de* from the oral register of BP in the 1900s

The search was limited to samples from the most recent century so that the data would reflect the most modern possible trends for BP. Duarte (1993) showed a decrease in null subjects from 1918-1992 of 25% to 75%. The present study is synchronic, examining the 1900s as a single unit. It is possible that the results of the present study are affected by the changes that Duarte observed over the past century. It is important to note, however, that the majority of the corpus samples from the 1900s are from the 1980s on, with the exception of the fiction register. Separate searches were performed for each of the four registers so that the results could be compared across registers in order to better answer the primary questions of this study.

After the specific search parameters had been set, the corpus showed the raw numbers for all the samples of the search item across the different sections of the corpus as shown in Figure 3-2.

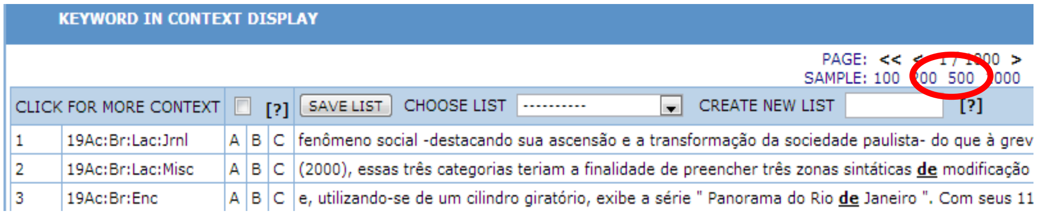


SEE CONTEXT: CLICK ON WORD (ALL SECTIONS), NUMBER (ONE SECTION), OR [CONTEXT] (SELECT) [HELP...]

	☐	CONTEXT	ALL ☐	1300s ☐	1400s ☐	1500s ☐	1600s ☐	1700s ☐	1800s ☐	1900s ☑	PORT ☐	BRAZ ☑	ACAD ☑
1	☑	DE	1830271	38505	92084	163218	121447	71511	362726	980780	494946	485834	304919

Figure 3-2 Selecting the correct section within the corpus

At this point, I selected the specific section of search results to be reviewed by selecting the number below the desired section, as demonstrated in Figure 3-2, where the results for *de* from the academic portion of the corpus have been selected. By selecting this section, I produced a comprehensive list of all instances of the search item in that section of the corpus, as illustrated in Figure 3-3.



1	19Ac:Br:Lac:Jrnl	A	B	C	fenômeno social -destacando sua ascensão e a transformação da sociedade paulista- do que à grev
2	19Ac:Br:Lac:Misc	A	B	C	(2000), essas três categorias teriam a finalidade de preencher três zonas sintáticas **de** modificação
3	19Ac:Br:Enc	A	B	C	e, utilizando-se de um cilindro giratório, exibe a série " Panorama do Rio **de** Janeiro ". Com seus 11

Figure 3-3 Creating a randomized list of samples within the corpus

After the corpus had produced the comprehensive list of all instances of the search item found within the specified section of the corpus, I was able to produce a randomized sample of language samples containing the search item. As shown in Figure 3-3, at the top right corner of the window, there is a sample option which allowed me to view a random list containing 100, 200, 500, or 1000 language samples. Figure 3-3 also demonstrates how the CDP showed the search items within context, allowing me to "click for more context" as shown on the left side of the page. It also gave specific information for each sample, including century, register, dialect, and source (the first sample in Figure 3-3 is labeled 19Ac:Br:Lac:Jrnl meaning that it was from the 1900s, in the academic register, in the Brazilian dialect, from an academic journal). As the purpose of this study was to answer more general questions about BP, the random sample option was used to produce samples representing a greater cross section of the language.

3.3 Quantity and type of samples collected

The corpus was set to produce a random list 1000 samples including *de* for each of the four registers (academic, spoken, fiction, and newspaper). Each sample produced by the corpus contains the nuclear word (*de)* with 180 to 200 surrounding words of context in the expanded-context view. From this expanded context, only independent, transitive sentences in main clauses were collected. Embedded clauses were not collected. Samples that had direct object pronouns and reflexive pronouns were not collected, as these types of pronouns are clitics in BP and behave differently than normal arguments. They cliticize to the verb and can thus appear in positions that are often not available to free morphemes (Duarte , 1989; Pagotto, 1992; 1993; Martins, 1994; Barbosa, 2000; Azevedo, 2002). While fascinating, clitic position is a topic for another project and was not examined here. The exclusion of clitics meant that with regards to covert objects, the present study only examined transitive clauses that had either overt objects that weren't pronominal, or that had covert objects.

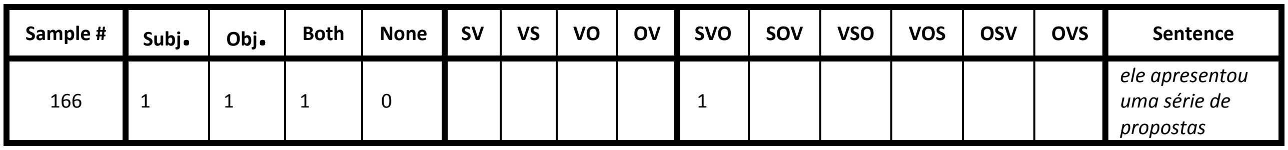
For each of the four registers, the first 250 transitive independent clauses were collected, creating a general list of 1000 transitive clauses. It was important that the samples be independent clauses, because embedded clauses can behave differently than main clauses in some languages, and it would be more difficult to isolate the features that the present study is designed to examine. For example, in English questions, main clauses show subject-verb inversion where embedded clauses do not (Azar, 2006).

3.4 Classifying the samples

After the searches were performed, the random sample sentences produced were categorized and recorded. Figure 3-4 illustrates this process as it was performed for the sentence *ele apresentou uma série de propostas* 'he presented a series of propositions' as seen in (1) taken from the newspaper register of the CDP:

(1)	***ele***	***apresent-ou***	***uma***	***série***	***de***	***propostas.***
	he	present-PST.3SG	a	series	of	propositions

‘He presented a series of propositions.’



Sample #	Subj.	Obj.	Both	None	SV	VS	VO	OV	SVO	SOV	VSO	VOS	OSV	OVS	Sentence
166	1	1	1	0					1						*ele apresentou uma série de propostas*

Figure 3-4 Classifying individual samples from the corpus

Each sentence was copied into the far-right column of the table. If there was an overt subject, the number 1 was placed in the “subject” column (as demonstrated by Figure 3-4). If not, 0 was placed in the column. The process was repeated for the “object” column. If both subject and object were present, a 1 was placed in the “both” column. If neither were overt, a 1 was placed in the “none” column. While it may appear redundant to have additional columns for “both” and “none”, this greatly facilitated the sorting of the data for the word-order analysis, as some null-subject clauses had overt objects, and some covert-object clauses had overt subjects. The creation of these columns meant that I could easily use a sort function to call up a specific type of sample in only one step. The next four columns were used to indicate the word order if either the subject or object was missing, with a 1 being placed in the appropriate column. The next six columns function much the same way, but are used in instances where the sentence has both an overt subject and object. A 1 was placed in the column indicating the correct word order for the sample sentence, in this case SVO (subject verb object). This format allowed for simple statistical analysis to be performed examining several key points of interest to this study to be further discussed in the next chapter:

1. The prevalence of the covert subject for each register
2. The prevalence of the covert object for each register
3. The frequency of each word order variation in sentences with overt subjects and objects for each register

3.5 Determining *topic* and *focus* in the samples:

The fifth main research question of the present study inquires as to the reasons behind the different variations observed in BP word order. As discussed in the literature review, Kato and Raposo (1996), and Costa (2000) discussed the discourse function of the different arguments as a possible motivation for word-order variation in Portuguese. They specifically mentioned the role of *topic* and *focus* within the sentence. Costa observed the following pattern for European Portuguese:

1. Preverbal definite subjects are old information.
2. Preverbal indefinite subjects are old information.
3. Postverbal subjects must be new information:
    a. If they precede the object, the object is also new information.
    b. If the object is not new information, the subject follows it (Costa, 2000:106).

Samples retrieved from the corpus were analyzed in order to determine whether or not the position of the topic (old information) and the focus (new information) followed the same pattern in BP that Costa observed in EP. In order to determine the position of the topic and the focus, I used the definitions established by Costa (p. 103-104):

1. In a questions-answer pair, a focused constituent in the answer replaces the *wh*-word in the question.
2. A topic is information already referred to in the discourse or a subpart of a given referent.

In addition to the definition provided by Costa, in the present study, a focused constituent was one that contained new information. New information was considered to be any information that did not precede the clause in question within the given context. The expanded context of each of the collected BP samples was examined to determine whether or not the arguments within the independent, transitive clause constituted new information (focus), or if they were referring to something previously mentioned (topic). If something was not mentioned previously in the context, it was labeled focus. If it was mentioned, it was labeled topic. Costa (2000) showed an example where an argument constituted old information, but wasn't the topic, as recreated in (17a) of chapter 2 in the present study. This was a *ditransitive* sentence (containing a subject, direct object, and indirect object). The present study examined *transitive* sentences with only one object. This helped control for the possibility of there being more than one argument that could be topical in the sample clauses.

As the context was very limited, it is possible that something determined to be focus in the available context could have been mentioned in the clause immediately preceding the context given by the corpus. Therefore, it was only possible to tentatively conclude that things were focus. Figure 3-5 is provided, because it shows the type of contextual information provided by the corpus, as well as the contextual examination of the clause glossed in (2) which was taken from the academic register of the CDP. This context is important, because it contains the preceding sentences and clauses which are important for determining the newness of the information provided by the different components of the analyzed clause:

(2)		S:Topic	V		O	
	a	***fisiologia***	***investig-a***	***os***	***mecanismos***	***de***
	the	physiology	investigate-3SG.PRS	the	mechanisms	of
	funcionamento		***do***	***organismo.***		
	function		of.the	organism		

‘Physiology investigates the organism’s functional mechanisms.’

; ao grande Filo Arthropoda pertencem os quelicerados como aranhas e escorpiões, os crustáceos como o camarão e o siri, e os insetos;, bem como explicar sua origem. Dessa forma, a zoologia acaba por relacionar-se a outros campos de estudo$_i$: **a fisiologia$_i$ por exemplo, investiga os mecanismos de funcionamento do organismo**; a paleontologia traz à tona o conhecimento das origens dos seres vivos hoje viventes; a citologia pode caracterizar os organismos ao nível celular, trazendo esclarecimentos em relação àquilo que o estudioso observa

Figure 3-5 Determining *topic* and *focus* of samples

The clause glossed in (2) is an example of an SVO sentence, in which the subject *a fisiologia* (physiology) is the topic of the sentence. This is determined by the fact that it is not new information, as it refers to something mentioned in the preceding sentence as shown in (3):

(3) a.	***dessa***	***forma,***	***a***	***zoologia***	***acab-a***	***por***	***relacionar-se***	***a***	***outros***
	of.this	form	the	zoology	finish-3SG.PRS	by	relate-REFL	to	other
	campos	***de***	***estudo$_i$.***						
	fields	of	study						

‘In this way, zoology ends up being related to other fields of study$_i$:’

b. **a *fisiologia*$_i$ *por exemplo, investiga os mecanismos de***
the physiology for example investigate-3SG.PRS the mechanisms of
funcionamento do organismo.
function of.the organism
‘Physiology$_i$, for example, investigates the organisms functional mechanisms.’

A fisiologia in (3b) is an example of one of the previously mentioned ‘other fields of study’ in (3a) that are related to zoology; therefore, it is “a subpart of a given referent” (Costa, 2000:104), and doesn’t constitute new information. For this reason, it is classified as the topic of the clause. This sentence is an example where the preverbal, definite subject that constitutes old information, meaning that for this sample, Costas EP word-order observations would hold true for BP as well.

3.6 Consultant

As I am not a native speaker of BP, I obtained approval from the Institutional Review Board to work with a consultant who is a native speaker. The consultant was from São Paulo, Brazil and had received a BA in Portuguese from a Brazilian university. At the time that this study was performed, the consultant was a MA candidate at Brigham Young University.

Due to the fact that there are several word order possibilities, and both subjects and objects can be dropped in BP, the consultant provided invaluable grammaticality judgments in situations where it would have been very difficult for a non-native speaker to classify a language sample, as illustrated by the sentence in (4) taken from the fiction register of the CDP:

(4) ***só val-em os dias idos***
only to.be.worth-PRS.3PL the.M.PL days gone
‘(They) are only worth the days past.’ OR ‘Only the past days are worth (it).’

Sentence (4) is either a transitive sentence with a covert subject, which would be classified as a VO sentence, or it is a transitive sentence with a covert object and a post-verbal subject, classifying it as a VS sentence. Both are grammatically possible, as the verb would be in agreement with either option as far as tense and person are concerned. In this situation, it fell to the consultant to determine what the correct classifications would be based on the context provided, and as it turns out, the second option (the covert object) was the correct one. The consultant also provided grammaticality judgments on some of the corpus sentences which exhibited more uncommon structures.

3.7 Statistical analysis

To determine whether or not the variations between registers and frequencies were significant, chi-squared statistical analyses were performed as described by Weisstein (1999). The chi-squared statistic is useful for comparing the frequencies of different categories to determine whether or not the differences could be produced by chance. In this case, the goal was to determine whether or not null subjects, covert objects, and word-order variations have different frequencies depending on register. If the chi-squared analyses retrieved statistically significant results, it means that the variations were, in fact, related to register. If the statistical results were not significant according to the chi-squared statistic, it meant that the variations could have happened by chance—not systematic, but random variability (Welkowitz, Cohen, & Ewen, 2006).

4 Results and Statistical Analysis

This chapter will present the word order data retrieved from the CDP as described in the third chapter of the present study in order to answer the primary research questions. The first section of this chapter will present the results for the prevalence of null subjects in BP. The second section will show the prevalence of covert objects. The third section will examine each of the different word orders that were present among the collected samples.

4.1 Null Subjects in BP

The first question examined in the present study addresses the prevalence of null subjects in BP. The raw corpus data reflecting the presence of null subjects per register is shown in Table 4-1. The *% exhibiting NS* column shows the percent of the BP samples in each register that do not have a phonetically overt subject, for example, 21.6% of the samples in the academic register have a null subject.

Table 4-1: Prevalence of null subjects in BP by register

Register	*# of sentences with NS*	*% exhibiting NS*
Academic	54	21.6%
News	29	11.6%
Fiction	94	37.6%
Oral	117	46.8%
Combined	294	29.4%

As shown in Table 4-1, 29.4% of the samples collected from the corpus have a null subject. This appears to be quite a significant amount. Examples of null-subject clauses from each of the registers are shown in (1):

(1) a. Academic
S V O
Ø ***trav-ou*** ***relações intelectuais.***
NS lock-3SG.PST relations intellectual
'(He) locked intellectual relations...'

b. Oral
S V O
Ø ***esij-o*** ***respeito em relação aos horários.***
NS demand-1SG.PRS respect in relation to.the hours
'(I) demand respect with regards to the schedule.'

c. Fiction
S V O
Ø ***assist-i*** ***a ocupação alemã.***
NS see-1SG.PST the occupation German
'(I) saw the German occupation.'

c. Newspaper
S V O
Ø ***mostr-a*** ***a vida de um policial.***
NS show-3SG.PRS the life of a policeman
'(It) shows the life of a policeman.'

To investigate the statistical significance of the observed difference between registers regarding the prevalence of the null subject, a chi-square statistic was used. The results of the Pearson chi-square analysis indicate that the four registers of BP examined within the corpus are significantly different on whether or not the samples exhibited a null subject ($\chi^2 = 63.578$, $df = 3$, $N = 1000$, $p < .0001$). Oral BP was the register with the highest number of null subjects at 46.8%, followed by fiction (37.6%), academic (21.6%), and finally newspaper (11.6%).

For the corpus samples with null subjects and overt objects, there were two possible word orders: verb object and object verb (VO and OV). The corpus data showing the prevalence of each of these two possibilities for the different registers is shown in Table 4-2:

Table 4-2: Prevalence of word orders for null subject samples by register

Register	*VO*	*%*	*OV*	*%*
Academic	52	94.5%	3	5.5%
News	27	96.4%	1	3.6%
Fiction	97	99.0%	1	1.0%
Oral	106	99.1%	1	0.9%
Combined	282	97.9%	6	2.1%

The *%* columns in Table 4-2 list the percentages of the null-subject samples that have an overt object and exhibit the specific word order. For example, 94.5% of the null-subject samples in academic BP have the VO word order. The combined percentages of all of the registers demonstrate that 97.9% of the samples retrieved from the corpus have the VO word order, where only 2.1% have the OV word order. VO is by far the more preferred word order.

In order to determine the statistical significance for the observed difference between registers regarding the prevalence of the different word orders for null-subject samples, a chi-square statistic again was used. The Pearson chi-square results for both the VO and OV samples indicate that the results regarding the VO and OV word orders by register are not statistically significant ($\chi^2 = 0.083$, $df = 3$, $N = 288$, $p = 0.9938$ and $\chi^2 = 4.524$, $df = 3$, $N = 288$, $p = 0.2101$ respectively). What this means is that for null-subject samples, the different registers behave in a very similar way with respect to word order, with an average of 97.9% of relevant samples following the VO pattern. It is important to note that for the OV order, the sample size was very limited; therefore, it is likely that the statistical analysis for this word order is not very meaningful (Welkowitz, Cohen, & Ewen, 2006).

4.2 Covert Objects in BP

The second question of importance to the present study deals with the occurrence of covert objects in BP. The raw corpus data showing the absence of overt objects per register is shown in Table 4-3. The *% exhibiting CO* column contains the percentage of samples in each register that have a covert object. For example, 4.4% of the corpus samples in the oral register have a covert object. It is interesting that only 2.3% of the samples have a covert object, where 29.4% have the null subject. Clearly null subjects are a lot more common than null objects in BP, a fact which chi-square analysis shows to be highly statistically significant ($\chi^2 = 225.804$, $df = 1$, $N = 311$, $p < .0001$).

Table 4-3: Prevalence of covert objects in BP by register

Register	*# of sentences with CO*	*% exhibiting CO*
Academic	1	0.4%
News	2	0.8%
Fiction	9	3.6%
Oral	11	4.4%
Combined	23	2.3%

An example of a CO clause from each register of the CDP is shown in (2):

(2) a. Academic
S V O
Toda estratificação implic-a Ø
every stratification implicate-3SG.PRS CO
'Every stratification implies (it).'

b. Oral
S V O
Eu abr-o Ø em Dezembro.
I open-1SG.PRS CO in December
'I open (it) in December.'

c. Fiction
S V O
Todo mundo confer-iu Ø
All world confirm-3SG.PST CO
'Everyone confirmed (it).'

c. Newspaper
O V S
Ø vai valer a determinação de cada
CO go.3SG.PRS to.be.worth the determination of each
'Everyone's determination will be worth (it).'

To investigate the statistical significance of the observed difference between registers regarding the frequency of the covert objects in BP, a chi-square statistic was implemented. The Pearson chi-square analysis indicates that the four registers of BP examined within the corpus are significantly different on whether or not their samples displayed a covert object ($\chi^2 = 13.000$, $df = 3$, $N = 1000$, $p = 0.0046$). Again, the oral register of BP had the highest number of covert objects at 4.4%, followed by fiction (3.6%, newspaper (0.8%), and finally academic BP (0.4%).

For the corpus samples with covert objects and overt subjects, there were two possible word orders: subject verb and verb subject (SV and VS). The corpus data showing the frequency of each of these two possibilities for the different registers is shown in Table 4-4:

Table 4-4: Prevalence of word orders for covert object samples in BP by register

Register	*SV*	*%*	*VS*	*%*
Academic	1	100.0%	0	0.0%
News	0	0.0%	2	100.0%
Fiction	4	66.7%	2	33.3%
Oral	6	100.0%	0	0.0%
Combined	11	73.3%	4	26.7%

As in previous tables, the % column tells what percent of the samples in each register belongs to the specified word-order category. For example, 100.0% of the six oral samples with a phonetically absent object and an overt subject had an SV word order. Although there were only 15 samples with an overt subject and a covert object, it was still possible to see that the SV order is much more prevalent than the VS order overall. The combined percentages of the different registers indicate that 73.3% of the covert object samples have the SV order, and only 26.7% have the VS order.

In order to determine the statistical significance of the observed variation between registers regarding the incidence of the different word orders for covert-object samples, a chi-square statistic again was applied. The Pearson chi-square results for both the SV and VS samples indicate that the results regarding the SV and VS word orders by register are not statistically significant ($\chi^2 = 2.175$, $df = 3$, $N = 15$, $p = 0.5369$ and $\chi^2 = 6.019$, $df = 3$, $N = 15$, $p = 0.1107$ respectively). It is important to note that chi-square calculations are only reliable when there are at least twenty cases across the four categories (in this case, registers) being compared. In this particular example, there were only 15 cases among those collected that had both a covert subject and an overt object; therefore, the *p* values indicating significance may not be meaningful. It is possible that a greater sample size could yield more meaningful results.

4.3 Word Order in BP

The third key research question that the present study examined with data from *the CDP* involves word order: of the six possible word-order variations for samples that have both subjects *and* objects that are phonetically overt (SVO, SOV, VSO, VOS, OSV, and OVS), which word orders actually occur in BP, and how does their prevalence vary by register? Interestingly, every possible word order was represented within the sample data, as illustrated by Table 4-5. It is important to note, however, that upon performing pragmatic analysis of the samples, the consultant determined that some of the samples in the more rare word orders presented by the corpus were incorrectly categorized due to punctuation errors, or there simply wasn't enough context to determine whether or not they were grammatical. In the case of SOV, all four of the samples were later judged to be incorrectly categorized, meaning that they belonged in some other category which could not be determined, or they were ungrammatical according to the consultant. Therefore, some of these findings presented here were not conclusive. This will be discussed in greater detail in chapter 5 of the present study.

Table 4-5: Prevalence of different word orders per register[1]

Register	*SVO*	*%*	*SOV*	*%*	*VSO*	*%*	*VOS*	*%*	*OSV*	*%*	*OVS*	*%*
Academic	185	95.4%	2	1.0%	0	0.0%	7	3.6%	1	0.5%	1	0.5%
News	212	96.4%	1	0.5%	0	0.0%	1	0.5%	0	0.0%	5	2.3%
Fiction	139	92.1%	1	0.7%	1	0.7%	0	0.0%	1	0.7%	5	3.3%
Oral	123	96.1%	0	0.0%	1	0.8%	0	0.0%	3	2.3%	2	1.6%
Combined	659	95.1%	4	0.6%	2	0.3%	8	1.2%	5	0.7%	13	1.9%[1]

The *%* column for each category shows the portion of samples collected per register with subjects and objects that have the specified word order. For example, 95.4% of the academic BP samples with both a subject and an object are SVO, as seen in Table 4-5. This is a *vast* majority, with hardly any occurrences of the other word orders. The combined percentages of the different registers show that 95.1% of the samples were SVO. OVS was the second most prevalent order at a much lower 1.9%, followed by VOS at 1.2%, OSV at 0.7%, SOV at 0.6% and finally VSO at only 0.3%. An example of each of the three most prevalent word orders is shown in (3):

(3) a. Fiction

S		V	O				
o	***navio***	***passava***	***uma***	***série***	***de***	***canaviais***	***verde-claros***
the	ship	pass-3SG.PST	a	series	of	reeds	green-light

‘The ship passed a series of light-green reeds’

b. Newspaper

V	O			S	
ganh-a	***importância***	***neste***	***cenário***	***o***	***leilão***
gain-3SG.PRS	importance	in.this	scenario	the	auction

‘The auction gains importance in this scenario.’

[1] The analysis performed in chapter 5 determined that the SOV, VSO, and OSV samples were either ungrammatical or incorrectly categorized due to errors in punctuation.

c. Fiction
O V S
outro tanto de elogios$_i$ teve o serviço de coquetel.
other as.much of praise had-3SG.PST the service of cocktail
'Just as much praise$_i$ received the cocktail service' (lit.).

In order to determine the statistical significance of the observed dissimilarity between registers with regard to the dominance of different word orders, a chi-square statistic was once again applied to the data. Pearson chi-square analysis was performed for each of the six word order possibilities. The results of the analysis show the statistical significance for register variation for each of the six possible word orders. They will be discussed one-by-one:

1. SVO

The chi-square results for this word order indicated that the difference between registers is not statistically significant ($\chi^2 = 0.199$, $df = 3$, $N = 693$, $p = 0.9778$). This p value indicates that the slight variations between registers for the percentage of SVO samples could be attributed to chance.

2. SOV

The chi-square results for this word order indicated that the difference between registers is not significant statistically ($\chi^2 = 1.507$, $df = 3$, $N = 693$, $p = 0.6807$). This p value indicates that the slight variations between registers for the percentage of SOV samples could be attributed to chance. Also, the low frequency (only four examples of SOV sentences across the four registers) means that the p value may not be meaningful.

3. VSO

The chi-square results for this word order demonstrated that the difference between registers again is not statistically significant ($\chi^2 = 3.002$, $df = 3$, $N = 693$, $p = 0.3913$). This *p* value indicates that the slight variations between registers for the percentage of SOV samples could be attributed to chance. It is important to remember that chi-square calculations are only reliable when there are at least twenty instances across the four registers. For this particular word order, only two samples were retrieved; therefore, the *p* values may not be very meaningful.

4. VOS

The chi-square results for the VOS word order provide evidence that the difference between registers is statistically significant ($\chi^2 = 14.273$, $df = 3$, $N = 693$, $p = 0.0026$). The academic register has the highest percentage of VOS samples at 3.6%, followed by newspaper (0.5%). Fiction and oral both had 0.0% occurrence of the VOS word order. There were less than twenty samples of VOS across the four registers (only eight VOS samples were retrieved from the corpus); therefore, the *p* values indicating significance may not be very reliable.

5. OSV

The chi-square results for this word order indicated that the difference between registers is not quite statistically significant ($\chi^2 = 6.378$, $df = 3$, $N = 693$, $p = 0.0946$). Here again, there were less than twenty instances of OSV across the four registers (only five OSV samples were collected from the corpus); therefore, it is possible that the *p* values are not reliable.

6. OVS

The chi-square results for this word order indicated again that the difference between registers is not significant statistically ($\chi^2 = 3.824$, $df = 3$, $N = 693$, $p = 0.2811$). Once more, this *p* value indicates that the slight variations between registers for the percentage of OVS samples could be attributed to chance. Also, the small sample size (only 13 examples of SOV sentences across the four registers) means that the *p* value may not be meaningful.

Another interesting characteristic of BP presented by the data is the rare (only seven samples out of 1000) but present possibility of neither subject nor object appearing overtly in the independent clause. Obviously there can be no word order variation when only the verb is phonetically present, but it is interesting to examine. Table 4-6 shows the number of corpus tokens per register that have both subject and object, as well as the corpus samples that have neither:

Table 4-6: Samples with both explicit subject and object, and samples with neither

Register	*Both*	*%*	*Neither*	*%*
Academic	194	77.6%	0	0.0%
News	220	88.0%	1	0.4%
Fiction	151	60.4%	0	0.0%
Oral	128	51.2%	6	2.4%
Combined	693	69.3%	7	0.7%

In Table 4-6, the *%* columns indicate what percent of the samples collected from each register express the *both* or *neither* property. For example, for the oral register, 2.4% of the collected BP samples have neither overt subjects nor objects. The combined totals of the different registers show that 69.3% of samples had both subjects and objects, where only 0.7% had neither. Example (4) shows a clause from the CDP with neither subject nor object expressed overtly:

(4) a. Oral

S	V	O
Ø	***aceit-ei***	**Ø.**
NS	accept-1SG.PST	CO

‘(I) accepted (it).’

To determine the statistical significance of the observed dissimilarity between registers regarding the occurrence of BP samples with both subject and object, as well as the prevalence of samples with neither, a chi-square statistic was used. The Pearson chi-square results indicate that the variation between the four registers of BP examined within the corpus is statistically significant for the *both* category ($\chi^2 = 29.776$, $df = 3$, $N = 1000$, $p < 0.0001$), and for the *neither* category ($\chi^2 = 14.143$, $df = 3$, $N = 1000$, $p = 0.0027$).

The present study found that newspaper BP was the register with the highest number of samples where both subject and object were phonetically present (88.0%), followed by academic (77.6%), fiction (60.4%), and finally oral (51.2%), whereas oral had the highest number of samples where neither were phonologically present (2.4%), followed by newspaper (0.4%), with only one sample, and fiction and academic each having no samples with neither subjects nor objects overtly expressed. Once again, it is important to note that with sample sizes as small as those shown in the “neither” category of Table 4-12, the chi-square calculations are not reliable, and the *p* value may not meaningfully or reliably represent the statistical significance of differences found between registers.

5 Discussion of Results

This chapter will discuss the results described in chapter 4. The sections of this chapter will be organized according to the primary research questions of the present study. The first section will examine null subjects, their prevalence, how they're influenced by register, and possible explanations for their occurrence. The second section will follow much the same format as the first, but it will address covert objects. The third section will discuss the frequency of all of the word orders observed in the corpus samples, how they differed by register, and how variations might be explained pragmatically. The fourth section will deal with the corpus itself, and whether or not it was an effective tool for collecting and analyzing this type of data.

5.1 Null Subjects in BP

The first question addressed in the present study tackles the prevalence of null subjects in BP. Overall, null subjects are quite common in BP as demonstrated by the corpus data. Across the four registers examined, 29.4% of the 1000 BP samples collected had null subjects. This provides a clear picture of null-subject behavior, answering the first question.

The fourth question inquired as to whether or not the frequency of null subjects might be influenced by register. Again, the corpus data provides a clear answer to this question, as demonstrated in Figure 5-1:

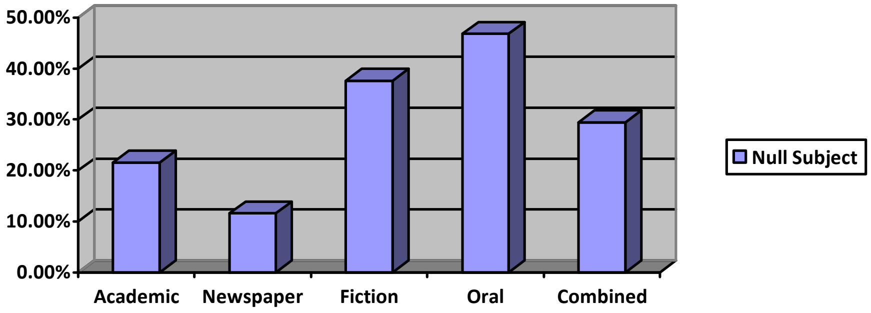


Figure 5-1 Frequency of null subjects by register

Statistical analyses of this variation showed the results to be very significant, meaning that register has a tremendous affect on the rate of null-subject pro-drop in BP. This means that pro-drop is not as prevalent in all registers, and that it is important to examine all of the registers before making any claims about the general pro-drop behavior of BP.

It is interesting that the oral register had the highest occurrence of null subjects, at 46.8%, meaning that almost half of the collected samples did not have a subject that was phonologically present. Newspaper, on the other hand, had significantly fewer examples with null subjects (11.6%). It is difficult to compare these results with those found by Duarte (1993), as he performed a diachronic study examining a decrease in the frequency of null subjects in the oral register of BP over the past century. The present study was more of a synchronic snapshot of null-subject behavior over the 1900s as a single unit. Duarte did, however, find an average null-subject frequency of 47.6% for oral BP in the 1900s. Duarte found an even higher rate of pro-drop in the 1800s, at 80% in 1845 and 77% in 1882 (Duarte, 1993: 112)(see Figure 2-1 in chapter 2 of the present study). Barbosa, Duarte and Kato (2005) found that 44% of BP samples taken from magazine interviews had null subjects (Barbosa, Duarte, & Kato, 2005:13)(see Table 2-1 in chapter 2 of the present study). Both of these results are remarkably similar to the present study’s null-subject findings of 46.8% for oral BP in the same period. The present study is valuable in that it provides contested results for null subjects in the oral register, but it also shows their prevalence in three additional registers of BP that were not examined by the literature. This is an important thing to examine, because the data collected in the present study show that pro-drop is much less frequent in written registers (oral BP was the register with the highest number of null subjects at 46.8%, followed by fiction (37.6%), academic (21.6%), and finally newspaper (11.6%)).

Why would the registers behave differently with regards to pro-drop? It is interesting to note that the more formal registers (academic and newspaper) have much lower frequencies for null subjects. It is possible that pro-drop is seen as casual, or inexact, much the same way that many reduced or contracted forms are viewed in English. For example, it would be inappropriate to write *he’d* (he would) or *gonna* in an academic paper, but these are quite common in the spoken registers of English (Azar, 2006). If pro-drop were perceived as informal, it could explain why it is common in the oral register, and even fiction (which can be less rigid and even tries to imitate the oral register at times), but less prevalent in the more rigid and formal registers of academic and newspaper reporting. It is also possible that the higher frequency of null subjects in oral and fiction is due to the large quantity of shared information between the speakers. Both participants in the conversation have a lot of background and contextual knowledge relating to the topic of the conversation, reducing the need for overt subjects. This has been described in English (Biber et al., 2002; Longman, 1999)The cause of variation among registers would be a fascinating question for additional study.

With regards to the fifth research question (why do null subjects occur in BP), the *topic/focus* analysis may provide part of the answer, as illustrated by the clause glossed in (1a).

This clause is one of the corpus samples collected from the academic register. (1a) shows the sentence immediately preceding the target sentence (1b) within the expanded context of the corpus. The verbs coindexed (i) both have a null subject, but were determined by the consultant to refer to the same subject:

(1) a. **posteriormente Ø$_i$** ***atu-ou como conselheiro do estado.***
afterwards NS act-3SG.PST as counselor of.the state
'Afterwards (He)$_i$ acted as counselor of the state.'

b. S:Topic V O:Focus
Ø$_i$ ***trav-ou relações intelectuais.***
NS lock-3SG.PST relations intellectual
'(He)$_i$ locked intellectual relations...'

The null subject of the sample in (1b) was analyzed as the topic, as it does not contain new information. This was slightly difficult to determine, as only part of the context is available in the corpus. The actual referent is never stated within the sample presented in the corpus; it was cut off, but the consultant determined that all of the preceding sentences were referring to the same null subject, as demonstrated by (1a) where the subject, while not phonologically present, was determined to be the same as that of the target sentence (1b). Therefore, the null subject does not contain new information and may be omitted as it can be inferred pragmatically as described by Downing and Noonan (1995) and Haspelmath (2001).

Another factor which may contribute to the frequency of dropped subjects is more grammatical in nature, as illustrated by (2), another academic sample:

(2) ***Ø pod-emos situar historicamente este nascimento.***
NS can-1PL.PRS situate historically this birth
'(We) can situate this birth historically'

The morphology of the verb *podemos* ‘we can’ contains all of the information necessary to tell us that this clause has a first-person, plural subject, and that the sentence is describing the present tense. This verb agreement in BP cross-refers with the verbs’ subject, making an overt subject unnecessary, as described by Haspelmath (2001:1500). Either the verb agreement, or the topical nature of the subject, or some combination of both could account for the option of using a null subject in BP, but these features still don’t explain why the null subject was more prevalent in the oral register than in written registers.

Previously, the possibility that null subjects were seen as informal was presented. It may also be possible that the frequency of first and second-person pronouns influences the prevalence of pro-drop, as these are more likely to be expressed with null subjects. The idea here is that if the subject is *the man* as in the sentence *the man went to the store*, you can’t have a null subject if *the man* isn’t previously mentioned in the context; whereas, if the subject is *I* or *you*, it can be expressed using a null subject because of verb agreement. Therefore, if a register has a higher frequency of first or second-person subjects than third-person subjects, it could potentially have more pro-drop sentences. Also, as previously mentioned, the high quantity of shared information between the two speakers makes overt subjects unnecessary at times (Biber et al., 2002; Longman, 1999).

To determine if this was a possibility in BP, the corpus samples were reexamined in the oral and newspaper registers. In the newspaper register, only 3.3% of the samples contained an overt or null subject that was first or second person. In the oral register, 56.7% of the corpus samples had overt or null subjects that were first or second person. This dramatic difference in the frequency of subject types could be evidence that the type of subject (first, second, or third-person) is one of the variables that affect the frequency with which null subjects occur, since the oral register has a much higher frequency of first and second-person subjects, and also a much higher occurrence of null subjects.

5.2 Covert Objects in BP

The second question of importance to the present study deals with the occurrence of covert objects in BP. Overall, covert objects are not as common as null subjects in BP, but they do occur as demonstrated by the corpus data. Across the four registers examined, 2.3% or the 1000 BP samples collected had covert objects (23 samples). Although the number of samples is small, the clauses exhibiting covert objects are not speech errors. According to the native-speaker consultant, they are grammatical within the context. An example from the oral register is included in (3), and the context will be explained later in this section:

(3)	***nós***	***nem***	***cham-amos***	***Ø***	***de***	***meditação.***
	we	NEG	call-1PL.PRS	CO	of	meditation

‘We don’t even call (it) meditation.’

According to the consultant, the sample in (3), and the others like it were not strange or ungrammatical; therefore, the corpus data collected successfully illustrated that covert objects do occur in BP, and it provided a clear measurement of their frequency, answering the second primary question of the present study. It is interesting to note that the frequency of covert objects measured in the present study is much lower than that reported by Cyrino (79.1%), but she included sentences with clitics, counting them as covert objects, where the present study did not collect clitic samples (Cyrino, 1993; 1994).

The fourth question inquired as to whether the presence of covert objects might be influenced by register. Again, the corpus data provides a clear answer to this question, as demonstrated in Figure 5-3:

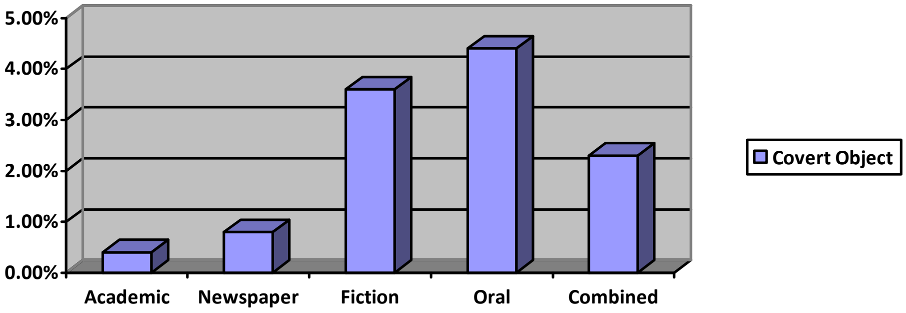


Figure 5-2 Frequency of covert objects by register

Statistical analyses of this variation showed the results to be very significant, meaning that register has a strong affect on the rate of covert object pro-drop in BP. This means that, here again, different registers do not exhibit pro-drop in the same way, and that it is important to examine all of the registers before making any claims about the general pro-drop behavior of BP.

As with the null subjects, the oral register of BP had the highest number of covert objects at 4.4%, followed by fiction (3.6%), newspaper (0.8%), and finally academic BP (0.4%). The present study was not concerned merely with the existence of covert objects, but also with possible reasons for their occurrence. Different from null subjects, there is no morphology that is cross-referential with the covert objects in BP, so the explanation must be elsewhere. An analysis of the topic and focus of the samples provided a possible solution. The oral sample from (3) is presented here again in (4b), glossed and with the topic labeled. The sentence that immediately preceded the target sentence (4b) in the expanded context provided by the corpus is also glossed (4a) with all of the co-referential words indexed (i):

(4) a. ***é*** ***que a base da prática***$_i$ ***é o Zazen.***
be.3SG.PRS that the base of.the practice be.3SG.PRS the Zazen
'That is, the base of the practice$_i$ is Zazen.'

b. S V O:Topic
nós nem cham-amos Ø$_i$ ***de meditação.***
we NEG call-1PL.PRS CO of meditation
'We don't even call (it)$_i$ meditation.'

The covert object of the sample in (4b) was analyzed as the topic, as it does not contain new information. The topic of the overall context is 'the practice' of Buddhism, and the consultant determined that the missing object referred to this same topic; therefore it was not new information and could be omitted. Just like the null subjects, these covert objects can be inferred pragmatically as described by Haspelmath(2001). The topical nature of the omitted objects accounted for all of the covert-object samples for which there was sufficient context (see section 4 of this chapter), providing an answer to the fifth research question.

Another interesting characteristic of BP related to covert objects and presented by the data, is the rare (only seven samples out of 1000, or 0.7%) but present possibility of neither subject nor object appearing overtly in the independent clause. This variation is grammatical within certain contexts, determined by the consult. An example taken from the oral register of the CDP is shown in (5):

(5) S V O:Topic
Ø ***aceit-ei*** **Ø.**
NS accept-1SG.PST CO
'(I) accepted (it).'

The corpus samples attest that it is possible for null-subject, covert-object sentences to exist in BP, and the consultant confirmed that they are grammatical contextually.

The statistical analysis performed in chapter 4 determined that register does affect the frequency of these null-subject, covert-object forms, as illustrated in Figure 5-6:

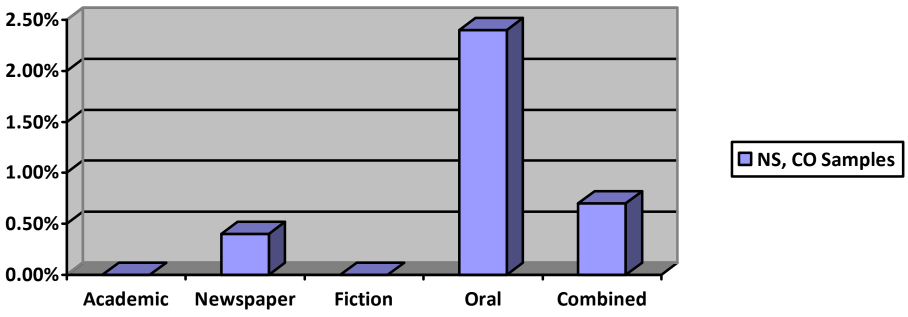


Figure 5-3 Frequency of null-subject, covert-object samples by register

As shown in Table 5-6, the oral register had the highest number of samples where neither were phonologically present (2.4%), followed by newspaper (0.4%), with only one sample, and fiction and academic each having no samples with neither subjects nor objects overtly expressed. Once again, it is important to note that with sample sizes as small as these, the statistical analysis may not be meaningful . It was also interesting that in every case, the covert object would have been third-person.

It is probable that the explanation for this variation is the same as that of the null-subject and covert-object variations, as this is probably a combination of the two, meaning that in samples where both are covert, they are both old information. In (5) the verb morphology makes it clear that the subject is the speaker, making an overt subject unnecessary. Here again, this agrees with Haspelmath's claims (2001: 1500). The covert object has been analyzed as the topic of the sentence. The consultant identified the object as the *simplificação do sistema* (simplification of the system) which is mentioned in the previous sentence, making the would-be object old information, and pragmatically deducible, as described by Chomsky (1981). The sentence could be translated as “I (the speaker) accepted the simplification of the system”.

5.3 Word Orders in BP

The third primary research question of the present study inquired as to which word orders actually occur in BP. In this section, each of the observed word orders will be described individually, with regards to their prominence, their variation across registers, and finally, possible causes.

5.3.1 Null subject variations

For the corpus samples with null subjects and overt objects, there were two possible word orders: verb object and object verb (VO and OV). The VO word order was the much more common of the two, making up 97.9% of the null-subject samples. The OV order obviously was much rarer, accounting for only 2.1% of the total null-subject samples retrieved from the corpus. Though rare, they were judged to be grammatical by the consultant. An example from the academic register is presented in (6):

(6)	***essa***	***característica***	***Ø***	***não***	***te-mos***	***na***	***rede***	***neural.***
	this	characteristic	NS	NEG	have-1PL.PRS	in.the	net	neural

‘(We) don’t have this characteristics in the neural net.’

According to the consultant, the sample in (6), and the others like it were not strange or ungrammatical; therefore, the corpus data collected successfully illustrated that both VO and OV word orders do occur in BP, and it provided a clear measurement of their frequency, answering the third primary question of the present study.

The fourth question inquired as to whether the frequency of the different word orders in the null-subject samples might be influenced by register. The statistical analyses determined that there was no statistical significance for this variation, although the small sample size (6 of the 1000 samples collected) means that the statistical analysis may not be meaningful . According to this analysis, register is not a big factor in determining how often these particular word orders occur. The registers all behave in about the same way with regards to this characteristic.

The present study also sought to determine the motivation for the varied word orders, especially as it relates to pragmatic features like topic. The VO word order seems to be the default for BP null-subject sentences. This is not surprising, as the most prevalent word order for BP is SVO, and it is probable that the subject has dropped out without affecting the word order. These findings are also in agreement with those presented by Costa for EP word order (2000), when he explained that post-verbal objects can contain either old or new information, meaning that it is normal for any type of object to occur postverbally.

The OV variation, on the other hand, is a little unusual. A contextual analysis identified the topic of the OV sentence. The academic sample from (6) is presented here again in (7b), glossed with the topic labeled. The sentence immediately preceding the target sentence within the expanded context provided by the corpus is glossed in (7a), with all of the co-referential words indexed (i):

(7) a. ***como Ø possu-i caraterísticas$_i$ simbolistas Ø pode***
as NS possess-3SG.PRS characteristics symbolistic NS can-3SG.PRS
ser adicionado à ferramenta.
be added to.the tool
'As (it) possesses symbolistic characteristics$_i$ (it) can be added to the tool.'

b.

O:Topic				V			
essa	***característica***	***Ø***	***não***	***te-mos***	***na***	***rede***	***neural.***
this	characteristic	NS	NEG	have-1PL.PRS	in.the	net	neural

‘(We) don’t have this characteristic$_i$ in the neural net.’

The object of the sample in (7b) was analyzed as the topic, as it does not contain new information. *Essa característica* ‘this characteristic’ refers to one of the ‘symbolistic characteristics’ *catacterísticas simpolistas* presented in (7a) which is the sentence immediately preceding the target sentence; therefore, it was not new information and could occur pre-verbally. For the six sentences with the OV order, the object was the topic in all of them. Here again, this is in line with Costa’s EP findings (2000) that objects can only occur pre-verbally when they are topicalized. This analysis suggests that the pragmatic features of topic and focus can explain these variations, where old information can occur before the verb, and new information must occur after.

5.3.2 Covert-object variations

For the corpus samples with covert objects and overt subjects, there were also two possible word orders: subject verb and verb subject (SV and VS). Both word orders were quite uncommon in the data, with only 11 SV samples and 4 VS samples among the 1000 samples collected. Though rare, they were judged to be grammatical by the consultant. Examples from the fiction register are presented in (8) (SV) and (9) (VS):

(8)

S		V	
todo	***mundo***	***confer-iu***	***Ø.***
all	world	confirm-3SG.PST	CO

‘Everyone confirmed (it).’

(9)

	V			S	
só	***val-em***	***Ø***	***os***	***dias***	***idos.***
only	to.be.worth-PRS.3PL	CO	the.M.PL	days	gone

'Only the days past are worth (it).'

According to the consultant, the samples in (8) and (9), and the others like them were not strange or ungrammatical; therefore, the corpus data collected successfully illustrated that both SV and VS word orders do occur in BP, and it provided a clear measurement of their frequency, answering the third primary question of the present study.

Overall, for the covert-subject samples, SV was more common than VS, making up 73.3% of the samples, with VS making up the remaining 26.7%. As far as register was concerned, there was no statistical significance to show that the registers behaved differently with regards to these particular variations, although, here again, there were only 15 total samples of covert-object clauses, meaning that the statistical significance may not be meaningful .

To determine whether or not topic and focus influenced the variation in word order for covert-object samples, the context of each clause was analyzed. The fiction sample from (8) is presented here again in (10b), glossed with topic. The expanded context for the sample is shown in (10a), with all of the co-referential words indexed (i):

(10) a.

quem$_i$	***duvid-asse***	***Ø***	***olh-asse.***
who	doubt-3SG.PST	CO	doubt-3SG.PST

'Whoever$_i$ doubted (it) might look.'

b.

S:Topic		V	
todo	***mundo$_i$***	***confer-iu***	***Ø.***
all	world	confirm-3SG.PST	CO

'Everyone$_i$ confirmed (it).'

The subject of the sample in (10b) was analyzed as the topic, as it does not contain new information. *Todo mundo* ‘everyone’ refers to the *quem* ‘whoever’ in the previous sentence; therefore it was not new information and could occur pre-verbally. Here again, this is in line with Costa’s EP findings (2000) that subjects can occur pre-verbally when they are topicalized, even if they are indefinite.

The VS samples were also analyzed to see if the location of the topic influenced word order. The fiction sample from (9) is presented here again in (11), glossed with the topic labeled:

(11)	V				S:Focus	
	só	***val-em***	**Ø**	***os***	***dias***	***idos.***
	only	to.be.worth-PRS.3PL	CO	the.M.PL	days	gone

‘Only the days past are worth (it).’

The subject of the sample in (11) was analyzed as the focus, as the “days gone” are not previously mentioned anywhere in the given context. It talks about ‘the days to come’ in the previous sentence, but “days gone” is a new idea; therefore it was new information.
Once again, this agrees with Costa’s EP findings (2000) that subjects only occur post-verbally when they contain new information. These findings suggest that the pragmatic features of topic and focus can explain SV and VS variations, where old information can occur before the verb, and new information must occur after. This held true for all four VS samples.

5.3.3 SVO variation

The SVO word order was very common in BP as demonstrated by the corpus data. This is not surprising, as Portuguese is classified as an SVO language (Azevedo, 2002). Across the four registers examined, 95.1% samples with both subject and object phonologically present exhibited this word order. This clearly illustrates the prevalent SVO behavior of BP, answering in part the third primary research question of the present study.

As to whether or not the frequency of the SVO word order was influenced by register, the statistical analysis performed in chapter 4 indicated that there is no significance to the SVO variation between the different registers examined within the corpus. This means that SVO has roughly the same level of occurrence across the board.

Regarding any pragmatic reasons for the prevalence of this order, the SVO samples were examined to determine whether or not they followed the same patterns that Costa observed in EP (2000). According to Costa, the SVO order with definite subjects may be used in either of two cases: the subject is familiar to the discourse participants but the object is not, or both subject *and* object are familiar (p. 104). This held true for BP, as illustrated in (12) (object is new information) and (13b) (both subject and object are familiar) which were taken from the fiction register of the CDP. The sentence that immediately preceded the target sentence (13b) is glossed in (13a) with coreferential constituents coindexed (i):

(12)

	S	V		O			
o	***navio***	***passava***	***uma***	***série***	***de***	***canaviais***	***verde-claros***
the	ship	pass-3SG.PST	a	series	of	reeds	green-light

‘The ship passed a series of light-green reeds’

‘The ship’ is the topic, and a definite subject (preceded by a definite article), but the object ‘light-green reeds’ is an indefinite object (preceded by an indefinite article), and is previously unmentioned in the text, making it new information.

(13) a. ***cada habitante de Sulidade tinh-a um concriz$_i$ que cant-ava***
each inhabitant of Sulidade have-3PL.PST a concriz that sing-3SG.PST
'Each inhabitant of Sulidade had a concriz$_i$ that sang.'

b. S V O
áte Cascofino consequ-iu o seu$_i$
even Cascofino obtain-3SG.PST the his
'Even Cascofino obtained his$_i$.'

'Cascofino' is the topic, and a definite subject (preceded by the definite article). It is also implied in the text that he is known to the participants (the use of the word 'even' implies that it would be surprising that Cascofino obtained his [concriz] based on some prior knowledge of Cascofino). The object 'his' is also known, as it is a possessive pronoun which has a coreferent in the preceding sentence (13a) (*concriz* is a type of yellow bird with a beautiful singing voice (Larousse, 2008)). The object is also preceded by a definite article. For these two contexts, the felicitous word orders from EP (Costa, 2000) were also strongly attested in BP.

Costa observed that the SVO order is also acceptable with indefinite subjects if they are not new information (p. 105). This held true for most of the indefinite samples analyzed ((14b) and (14a) which is the sentence immediately preceding (14b) in the expanded context, taken from the fiction register of the CDP):

(14) a. Ø$_i$ ***pag-avam bem e pontualmente.***
NS pay-3PL.PST well and punctually
'(they)$_i$ paid well and on time.'

b. S V O
alguns$_i$ porém quise-ram mais ainda.
some however want-3PL.PST more still
'Some$_i$, however, wanted more still.'

In (14b), the indefinite subject *alguns* 'some' is referring to a subset of the subjects of the previous sentences, according to the consultant, meaning that the indefinite subject is not new information. This agreed with Costa's observations. There was, however, one instance that did not agree, shown in (15) which was taken from the fiction register of the corpus:

(15)

	S			V		O
um	***rumor***	***de***	***falas***	***ench-ia***	***a***	***casa***
a	rumor	of	words	fill-3SG.PST	the	house

'A rumor of words filled the house.'

In (15), the indefinite "rumor of words" is not previously mentioned in any way, but rather it is the next event in a narrative sequence. Remembering that this register is fiction, this one sample may not actually illustrate a significant difference between BP and EP. In fiction, there is a lot of creative and poetic license that may allow for slight variations. It is also possible, here again, that the limited nature of the context available in the corpus is affecting the results. Just as Costa concluded for EP, from the preceding examples, it appears that preverbal subjects must constitute old or accessible information in BP (Costa, 2000).

5.3.4 VSO variation

Where SVO was a very common word order, the VSO word order was very rare in BP as demonstrated by the corpus data. Across the four registers examined, 0.3% of the samples with both subject and object phonologically present exhibited this word order, making it the least prevalent of the six logical word-order possibilities. Only two samples were collected from the corpus, and both appeared strange to the consultant. They are shown in (16):

(16) a. Fiction

V	S		O
*_**Tem**_	_**as**_	_**costas**_	_**sentença.**_
have.3SG.PRS	the	back	sentence

‘The back has a sentence.’

b. Oral

V	S		O
*_**Pens-o**_	_**eu**_	_**é**_	_**Paulo.**_
think.1SG.PRS	I	be.3SG.PRS	Paulo

‘Think I it is Paulo (lit.).’

(16a) probably contains some sort of error, as the verb *tem* is singular and doesn’t agree with the plural subject *as costas.* It is likely that the sentence was supposed to be *Tem NAS costas sentença* ‘(He) has a sentence on his back (lit.)’ meaning that he has been given a weighty sentence. This would be a null-subject sentence, but it was difficult to determine if this was correct. With regards to (16b), the consultant thought that this might actually be two sentences: *penso eu* and *é Paulo*. It is possible that the punctuation was missing. For this reason, the results for the VSO word order were inconclusive in this study.

As to whether or not the frequency of the VSO word order was influenced by register, the statistical analysis performed in chapter 4 indicated that there is no significance to the VSO variation between the different registers examined within the corpus. This means that all four of the registers approach this word order in a statistically similar way. Here again, the sample size was so small and the samples themselves so strange that the statistical analysis may not be meaningful.

Regarding any pragmatic reasons for the prevalence of this order, Costa explained that in EP, the VSO order is best when both subject and object are new in the discourse (p. 105). The present study does not have adequate data to determine if this holds true for BP.

5.3.5 VOS variation

Another present but uncommon word order in BP was VOS. As demonstrated by the corpus data, VOS made up 1.2% of the samples. These samples, while rare, were determined to be grammatical by the consultant. The sample from the newspaper register is shown in (17):

(17)	V	O				S
	ganh-a	***importância***	***neste***	***cenário***	***o***	***leilão***
	gain-3SG.PRS	importance	in.this	scenario	the auction	
	‘The auction gains importance in this scenario.’					

Regarding the VOS word order, the corpus data was effective in determining whether or not it was attested and how common it was.

The number of samples was very limited, so the statistical analysis performed in chapter 4 may not be entirely reliable, but it did show that VOS is one word order that is affected by register, as shown in Figure 5-14:

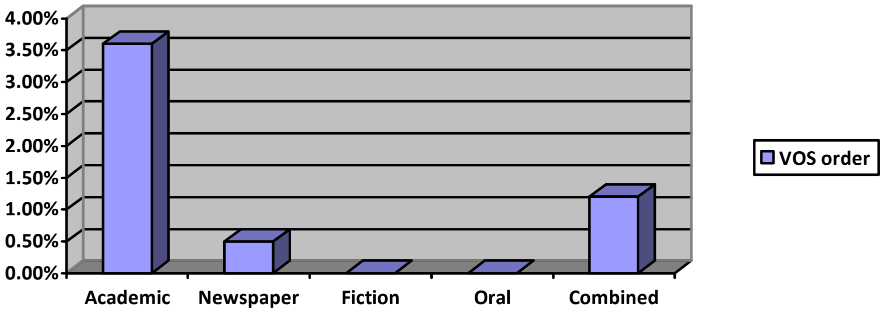


Figure 5-4 Frequency of covert objects by register

From Figure 5-4, it is very obvious that the VOS word order was much more common in Academic (3.5%), than in the other registers (0.5% for academic, and 0.0% for fiction and oral).

As with the VSO order, some of the samples in the academic register seemed a little strange to the consultant, which could be due to a lack of context, but the newspaper example in (16) was particularly clear.

In order to determine pragmatically why this word order occurs, the location of old and new information was examined. Costa (2000:105) observed that in EP, the VOS word order was possible if only the subject consisted of new information. In (18), the newspaper sample from (17) is once again presented with the components consisting of new information labeled as focus:

(18)	V	O:Topic				S:Focus
	ganh-a	***importância***	***neste***	***cenário***	***o***	***leilão***
	gain-3SG.PRS	importance	in.this	scenario	the auction	

‘The auction gains importance in this scenario.’

In (18) the subject *leilão* (auction) was analyzed as a *focus* element, as it was not previously mentioned in the text. The object on the other hand (*importância neste cenário* meaning “importance in this scenario”) is not new, in that all of the previous sentences are talking about the state of affairs, setting up the scenario for which caused the auction to gain importance. From this observation, it appears that Costa’s description of EP is consistent with the VOS behavior of BP, although more conclusive results would require a larger sample size.

5.3.6 OVS variation

Slightly more common than VOS, the OVS word order was also observed in the samples retrieved from the corpus. As demonstrated by the corpus data, OVS made up 1.9% of the samples. These samples were determined to be grammatical by the consultant, but they are very heavily dependent on context as most of them are direct responses to questions, either in interviews, or in fictitious dialogue. A sample from the fiction register is shown in (19):

(19) O V S
outro tanto de elogios$_i$ teve o serviço de coquetel.
other as.much of praise had-3SG.PST the service of cocktail
'Just as much praise$_i$ received the cocktail service' (lit.).

Regarding the OVS word order, the corpus data was effective in determining whether or not it was attested and how common it was. There was not a large enough data sample to determine significance of variation between registers with any degree of certainty, but it appears that there were no significant differences related to register.

In order to determine pragmatically why this word order occurs, the location of old and new information was examined. Costa (2000:105) observed that in EP, the VOS word order was possible if only the subject consisted of new information (the same context as the VOS word order). In (20b), the newspaper sample from (19) is once again presented with the components consisting of new information labeled as focus. (20a) contains the clause immediately preceding the target sentence as retrieved from the expanded context provided by the corpus:

(20) a. ...***unânimes nos elogios$_i$ à coleção.***
unanimous in.the praise of.the collection
'...[they were] unanimous in their praise$_i$ of the collection' (lit.).

b. O:Topic V S:Focus?
outro tanto de elogios$_i$ teve o serviço de coquetel.
other as.much of praise had-3SG.PST the service of cocktail
'Just as much praise$_i$ received the cocktail service' (lit.).

In (20b) the subject *o serviço de coquetel* 'the cocktail service' could be analyzed as a focus element. At least within the context available in the corpus, it is not specifically mentioned at any point previously. The object on the other hand (*outro tanto de elogios* or 'just as much praise' is not new. In the previous sentence, the article is talking about the praise given by critics to a different event. The target sentence is saying that the cocktail service received the same level of praise. From this observation, it appears that Costa's description of EP is consistent with the OVS behavior of BP, although more conclusive results would require a larger sample size. Costa (2000) states that both the VOS and OVS word orders are derived by object left-dislocation, which he describes as not surprising, since the object is old information (p. 106). The idea of word-order variations come from some sort of dislocation or clefting is not new, having been described for English and other languages as well (Downing & Noonan, 1995; Trujillo).

5.3.7 OSV variation

The OSV word order was also present among the corpus samples. As demonstrated by the corpus data, OSV made up only 0.7% of the samples. There were only five samples, and (like the VSO samples) they appeared strange to the consultant. These samples are shown in (21):

(21) a. Academic

O				S	V
****O mausoléu***	***de***	***Helicarnasso***	***quatro***	***escultores***	***trabalhar-am.***
the mausoleum	of	Helicarnassus	four	sculptors	work.3PL.PST

'The mausoleum of Helicarnassus four sculptors worked (lit.).'

b. Oral

O		S	V
****Uma***	***perícia***	***ele***	***diz.***
A	skill	he	say.3SG.PRS

'A skill, he says.'

c. Oral
O S V
****Quantos dias você tem.***
how.many days you have.3SG.PRS
‘How many days you have.’

d. Oral
O S V
****A educação doméstica eu ach-o.***
the education domestic I think-1SG.PRS
‘Domestic education I think (lit.).’

e. Fiction
O S V
****Onde? a voz respond-ia.***
where the voice respond-3SG.PST
‘”Where?” the voice responded.’

Example (21a) appears to have been punctuated incorrectly , causing them to be incorrectly identified. It is probable that the first part of the sentence *O mausoléu de Helicarnasso* was actually some sort of section heading that should have been separated by a line or by punctuation. Examples (21b) and (21d) appear to be sentence fragments where the interviewee didn’t respond with a complete sentence, and (21c) is an incorrectly punctuated question. Example (21e) is dialogue from a fictional interaction, which accounts for the strange word order. For these reasons, the results for the OSV word order were inconclusive in this study.

As to whether or not the frequency of the OSV word order was influenced by register, the statistical analysis performed in chapter 4 indicated that there is no significance to the OSV variation between the different registers examined within the corpus. This means that all four of the registers approach this word order in a statistically similar way. Here again, the sample size was so small and the samples themselves so strange that the statistical analysis may not be meaningful.

Regarding any pragmatic reasons for the prevalence of this order, Costa explained that in EP, the OSV order can occur when the object has been topicalized (p. 105). The present study does not have adequate data to determine if this holds true for BP.

5.3.8 SOV variation

The SOV word order, considered ungrammatical in EP by Costa (2000), initially seemed to occur in the corpus samples, if very rarely. As demonstrated by the corpus data, SOV made up 0.6% of the samples. While this was potentially interesting, a closer examination of these sentences, with the help of the consultant, determined that they were not good samples for the present study as the pre-verbal objects were clitics. An example from the CDP is shown in (22):

(22) a. Newspaper

S			O	V
A	***mesma***	***imprensa***	***se***	***encarreg-ou.***
the	same	press	REFL	charge-3SG.PST

‘The same press made itself responsible.’

These samples had initially been incorrectly categorized due to the fact that pre-verbal object pronouns (in the case of (22) a reflexive object pronoun) are not orthographically connected to the verb, but the literature clarified that they are still clitics (Azevedo, 2002; Barbosa , 2000). Therefore, these samples were thrown out. Statistical analysis also determined that there was no significance associated with this variation. A lack of adequate samples also made it impossible to examine pragmatic motives for the SOV word order.

5.4 The CDP

It is clear from the previous sections that BP exhibits some interesting variations when it comes to pro-drop and word order. A significant amount of empirical data was gathered and analyzed. Now, it is important to discuss the tool with which the data was obtained: the CDP. For the most part, this corpus was very useful for examining the principle research questions of the present study, but it wasn't perfect. This section will mention some of the benefits of using the corpus for this type of syntactic research, as well as some of the difficulties, as experienced in the present study.

First of all, using the CDP provided some definite advantages over more traditional methods for gathering syntactic data. For example, had this study relied entirely upon language samples provided by a native speaker consultant or consultants, it would have taken weeks, if not months to gather the same amount of relevant data. This methodology required a very specific type of clause, and it would be difficult to elicit only main clauses, with transitive verbs and no clitics, in an expanded context. With the corpus, it was relatively easy to collect 1000 of these language samples, making it possible to start creating a general picture of BP word-order behavior.

More than just providing a large *quantity* of data, the corpus provided large *variety* of data. Not only were the samples drawn from four different registers, but within each register there were dozens, if not hundreds or sources. This diversity means that the data are more representative of the language as a whole, whereas a study involving only a few consultants, or text from a handful of magazines could not make the same claims. Here again, the CDP was very useful in provided the variety of data needed to investigate the syntactic behavior of BP.

In addition to providing simple access to a large and varied data supply, the CDP solves some of the problems that Chomsky had identified with early corpus and empirical research as described in the literature review. One of Chomsky's largest criticisms of corpora of the era was that they were too small (Davies, 2008: 151). The CDP consists of 45 million words. While it is still one of the smaller corpora produced by Dr. Davies at Brigham Young University (the Corpus of Contemporary American English has 450 million words, and now the Global Web-Based English corpus has 1.9 *billion* words) (Davies, 2013), it is still in the tens-of-millions range, more than solving for Chomsky's size concerns.

Chomsky's initial arguments against early methods of data gathering had led to a fairly general adoption of a more theoretical method of syntactic analysis, relying heavily on the researchers own intuition and grammaticality judgments, as described in the literature review (Davies, 2008: 150). The corpus, on the other hand, serves as a standard for the language, providing actual data taken from real-world sources and contexts. the CDP was invaluable for this type of research, because there were no theoretical contexts or constructions. All of the clauses analyzed actually occurred in the language. This removed the responsibility of grammaticality judgment from the researcher's shoulders, allowing for the description, analysis, and documentation of the language as it is, not as the researcher thinks it might be. Using this corpus also meant that research could be perfomed on a language of which the primary investigator was not a native speaker, where most of the authors in the literature were native speakers of Portuguese.

While the benefits to using the CDP were invaluable for this word-order study, it wasn't without difficulties. Even though the corpus as a whole is very large, by the time it had been filtered by variety (BP vs. EP), time period (1900s), and register, the actual data was a much smaller subset. For example, for BP in the 1900s, the oral register only has one million words. For the present study, the overall size of the corpus did not truly present any difficulties, as only 250 samples were needed from each register, and the CDP provided more than enough data. Perhaps the size of the corpus could be a limiting factor for some other types of research.

Another issue involves the time periods that are available to be searched. This particular corpus is divided into sections consisting of one century each. It is impossible to look at a particular year or even a decade, therefore the conclusions of this study are only as specific as the twentieth century. Furthermore, I cannot make any observations specific to the last ten years. Languages are in a constant state of change, and it is possible that some of the results of the present study are obsolete.

Two of the biggest frustrations of this study are related to copyright laws, and therefore, they don't have an easy solution. First of all, I was limited to retrieving 200 expanded contexts per day due to copyright restrictions. Since I needed to filter through several thousand samples to find the 1000 that met the requirements of this project, this slowed the overall progress quite a bit. These restrictions are understandable, and not without legitimate reason. All things considered, 200 samples is still a significant amount of data collection for one day's work when compared with other methodologies (it is doubtful that doing fieldwork with a native speaker, for example, could produce this much data in this timeframe), but it is important to take these limits into account when planning this type of corpus research.

The other issue relating to copyright laws is the amount of expanded context that is available for each sample. As described in the sections above, there were quite a few instances where the context was not sufficient to determine pragmatic features like focus. This made it difficult to analyze some of the samples, and in some cases (especially with the rarer word orders which depend heavily on context) it made it difficult for my consultant to interpret the sentences. This meant that I was unable to verify if the clauses were full utterances or the product of punctuation errors in the transcription, etc. This wasn’t a problem for the word orders that had dozens of samples, but definitely limited my conclusions for those that had less than ten. There are ways to get around this limitation. While not super convenient, it is possible to chain several expanded contexts together by performing a specific search for the first few words of an expanded context and retrieving their expanded context and so on.

6 Conclusion

The aim of the present study was to use data retrieved from the CDP to examine pro-drop behavior in BP, compare the affect that topic and focus has on BP word order with that of EP word order, and also show how large corpora can be used for syntactic research. The first section of this chapter will present the conclusions of the present study with regards to the primary research questions of the present study. The second section will address the limitations of this study, and the third section will discuss possible directions for future research.

6.1 Answering the questions

This research took on the task of answering six main questions about syntactic behavior in BP:

1. How frequent are null subjects in BP?
2. How frequent are covert objects in BP?
3. Which of the possible word order variations actually occur in BP, and how frequent are they?
4. How are the frequencies of 1, 2, and 3 affected by register?
5. Why do null subjects, covert objects, and word order variations occur?
6. Is *the CDP* a viable source of data for syntactic and word order research?

With the current methodology, it was possible to answer each of these questions at least in part.

6.1.1 Null subjects and covert objects in BP

In chapter 2, the literature established BP as a pro-drop language, in that it can exhibit null subjects and null objects. Several studies measured to some extent the frequency of the null-subject feature in certain registers of BP (Duarte (1993) for the oral register, and Barbosa, Duarte, and Kato (2005) for written BP in magazines), but while these studies found compelling results, they were limited either by the size or nature of their corpora. Using the CDP, the present study was able to determine not only the overall rate of null subjects in BP (29.4% of sentences have null subjects), but it was also able to compare four different registers in a way that no other study had. This study also measured the occurrence of covert objects in BP, observing that 2.3% of BP sentences have this feature. These pro-drop results were clear, and successfully describe the tendencies of BP towards these interesting syntactic variations.

6.1.2 Word-order variations in BP

Not only did the literature discuss pro-drop for BP, but it also described word-order variations at great length. This is relevant to the pro-drop conversation, as one feature common to all null-subject language is subject-verb inversion (Chomsky, 1981; Rizzi, 1982). While BP is primarily an SVO language (Azevedo, 2002) other possible word orders have been discussed at great length in the literature. Linguists have done a lot to explain possible causes for word order variation (Costa, 2000; Kato & Raposo, 1996; Silva, 2001), but these researchers did very little to demonstrate how common (or uncommon) any of the possible variations were.

As with the pro-drop phenomenon, corpus data was able to answer the word-order question. In BP, 95.8% of independent, transitive clauses are SVO, (justifying the classification of BP as an SVO language). The corpus also demonstrated the existence of VOS and OVS variations that not only occurred, but were verified as contextually grammatical (not speech or transcription errors) by the native-speaker consultant. The corpus also produced instances of SOV, VSO, and OSV sentences, but for these variations, either there was not enough context provided to determine gramaticality, or there may have been some textual or speech error that caused them to be incorrectly categorized. In any case, for these last three possibilities, the findings of the present study were not conclusive.

6.1.3 The affects of register on variation

One advantage to using the CDP that other researchers did not have was the ability to examine the different registers of BP (academic, newspaper, fiction, and oral) (Davies & Ferreira, 2006; Davies, 2008; Davies, 2013). This meant that it was possible to determine whether or not pro-drop and word-order variations were consistent in the four main types of BP (chapters 4 and 5). Statistical analysis of the corpus data showed that the pro-drop feature is strongly affected by register, with the oral register having the highest frequency of both null subjects and covert objects.

Word-order variations, on the other hand, were not strongly affected by register at least within the limited sample size of the present study (1000 sentences). With the exception of the VOS word order, which was most common in academic and newspaper, statistical analysis of the corpus data showed that the registers did not vary significantly. It is important to remember that for these more rare variations, the number of samples was so small that the statistical analyses may be inaccurate regarding significance. This made some of the results inconclusive, but in the case of SVO, there were hundreds of samples, and the statistics determined that register did not significantly affect frequency.

6.1.4 Explaining variation

Having determined the existence and prevalence of both pro-drop and word-order variations in BP, it was important to answer the question regarding possible causes. For null subjects, the literature determined two possible explanations for the unnecessary nature of overt subjects: pragmatic inferability, and subject-verb agreement (Trujillo; Chomsky, 1981; Rizzi, 1982; Downing & Noonan, 1995; Longman, 1999; Haspelmath, 2001; Biber et al., 2002). The present study showed that both of these explanations work for null subjects in BP, as the null subjects in the corpus samples where either pragmatically inferable based on the context (old information) or the verb morphology made it obvious what the subject was, or both. Covert objects were also easily inferred from the context (all of the covert objects were analyzed as topic), but there is no BP verb morphology that references the object.

For word order, the literature proposed that variations were not random or context neutral, but depended on pragmatic features of *topic* and *focus* (Costa, 2000) (Kato & Raposo, 1996). Table 6-1 compares previous findings for BP and EP word order with those of the present study:

Table 6-1: BP and EP word order by topic and focus

Word Orders	*Present Study*	*Kato and Raposo (1996)*		*Costa (2000)*
	BP	*BP*	*EP*	*EP*
SVO	S:Topic V O:Topic/Focus	-	S:Topic V O:Focus	S:Topic V O:Topic/Focus
SOV	-	-	-	*S O V
VSO	-	-	-	V S:Focus O:Focus
VOS	V O:Topic S:Focus	-	-	V O:Topic S:Focus
OSV	-	O:Topic S:**FOCUS** V	-	O:Topic S V
OVS	O:Topic V S:Focus	O:**FOCUS** V S:Topic	O:Topic V S:Focus	O:Topic V S:Focus

Examining the expanded contextual information provided by the corpus, the present study determined that Costa's EP findings (2000) held true for BP as well, to the extent that there were samples to analyze:

1. Preverbal definite subjects are old information.
2. Preverbal indefinite subjects are old information.
3. Postverbal subjects must be new information:
    a. If they precede the object, the object is also new information.
    b. If the object is not new information, the subject follows it (Costa, 2000: 106).

These word-order guidelines adequately described the word orders found in the corpus samples. The one area where the findings of the present study differed from word-order behavior described by the literature was with the OVS word order. Kato and Raposo (1996) determined that this order was possible if the object was marked prosodically. The present study only found examples of topical objects in the preverbal position. It is important to note, however, that a written corpus without actual audio samples of speech cannot be used to determine if something was marked prosodically, so the findings of the present study only apply to a prosodically neutral context. This accounts for the difference between the present study and the literature. It appears from the present study that word order is not random. It is pragmatically governed.

6.1.5 Using the CDP

In chapter 2, there is a discussion of the historical debate of corpus linguistics (Davies, 2008; McEnery & Wilson, 2001). Davies (2008) established criteria for a type of corpus that both corrects problems associated with traditional empirical research, and answers criticisms leveled by proponents more theoretical or intuitive methods for syntactic research (p. 162):

1. Size: useful corpora typically contain tens of millions of words of text
2. Representativity: the best corpora will contain texts from a wide range of genres
3. Annotation: the texts will be lemmatized and will be tagged for part of speech
4. Architecture and interface: it will be possible to search by substring (for morphology), lemma and part of speech (for syntax), collocates and synonyms (for semantics), and frequency in different historical periods and registers (for lexical research and for stylistics and historical linguistics)

One of the goals of the present study was to determine if the CDP both meets these criteria and is effective for syntactic research. For 2, 3, and 4 on the previous list, the corpus is very well constructed. It was easy to search for the specific samples that were needed, and it was simple to examine the different registers BP. One possible problem is with the size of the corpus. As a whole, it is technically a large corpus by Davies' standards, but when filters are applied in order to examine a specific subset of the corpus, the number of samples is much smaller (only 1 million words in the oral register of BP for the 1900s). For the purposes of the present study, this was still plenty large, as only 250 samples were needed. The corpus is also limited in terms of searchable time periods. It was only divided up according to century, making it difficult to draw any conclusions about the most current state of the language, or about specific decades or diachronic trends.

The biggest difficulty with using the corpus for the present study was the limited nature of the expanded contextual information. It was necessary to examine the expanded context to do the pragmatic analysis of the samples, but due to copyright limitations, it was only possible to view 200 expanded samples per day, and the samples themselves were too short at times to determine topic and focus for the target clause. Clearly there is nothing to be done to change the copyright laws, but it is important to take these factors into account when designing a syntactic study that depends on this corpus. There are ways around these limitations. Further experimentation with the CDP has shown that it is possible to perform a specific search for the first few words of an expanded context, allowing the researcher to string two or more sections of expanded context together. For the most part, the CDP was extremely useful and easy to use, and the data allowed for many conclusions to be drawn.

6.2 Limitations

While the present study was successful in answering the key research questions, it wasn't without limitations. The biggest of these was the sample size. While initially 1000 clauses seemed to be quite a large data set, when it came time to run the numbers, some of the different features did not have enough samples to perform conclusive statistical analysis of significance. For several word orders, fewer than ten occurrences were documented within the data set. While this does indicate their existence, it was not enough to determine any significant affects for register. The limited sample size also made it difficult to determine topic, focus, and contextual grammaticality. This was due to the fact that the context for any sample was limited to a point, but with hundreds of samples it was possible to find many examples with sufficient context, whereas with smaller data sets the probability of getting a sample with sufficient context was much lower. Perhaps with a larger sample size it would have been possible to have more conclusive findings for SOV, VSO, and OSV, for example. With only 1000 samples analyzed, the conclusions of the present study might not reflect the overall word-order tendencies of BP, and could certainly be improved by a larger sample size.

6.3 Future work

Much was learned from the present study, and it could serve as a gateway for several future research projects. Probably the biggest thing demonstrated by this study was the usefulness of a large corpus, specifically the CDP, for performing syntactic research based on empirical data. This study focused specifically on pro-drop and word order variations in BP, but it would be possible to apply this methodology to studies examining any number of syntactic variations: clitic-placement, question formation, the prevalence of different pronoun types, etc. The present study made comparisons of different registers within BP, but it would be very easy to compare across dialects (BP and EP) or even across languages (BP and English, using the *Corpus of Contemporary American English* which uses the same architecture as the CDP (Davies, 2013). Using corpora to collect empirical data can confirm or clarify the findings of numerous syntactic studies.

As this study did expose some weaknesses associated with using the CDP, one future direction could be to improve the corpus in terms of size (although for the present study the CDP was capable of providing more than enough data) and searchable time periods, or even create a new Portuguese corpus incorporating the useful features of the old one, but much larger in size and scope. Just like English, Portuguese is a global language. There are significant Portuguese-speaking populations on every continent except Antarctica (Brazil, United States, Portugal, Mozambique, Angola, Guinea Bissau, Cape Verde, East Timor, Goa in India, Macao in China, and Japan, among others) (Azevedo, 2002). It would be interesting to have data for all of these dialects and regions so that greater comparisons could be made. The CDP is a powerful beginning, showing that it can be done, but there is much left to do.

References

Adger, D. (1995). Optionality and the Syntax/Discourse Interface. *Paper presented at the International Conference on Interfaces in Linguistics, November 1995.* Oporto.

Ambar, M. (1992). *Para uma Sintaxe da Inversão Sujeito Verbo em Português.* Lisbon: Ph. D Dissertation. University of Lisbon.

Azar, B. (2006). *Understanding and Using English Grammar, Third Edition.* Upper Saddle River, NJ, USA: Pearson and Longman.

Azevedo, M. M. (2002). *Portuguese: a linguistic introduction.* Cambridge: Cambridge University Press.

Barbosa, P. (2000). Clitics: A Window into the Null Subject Property. In J. Costa, *Portuguese Syntax* (pp. 31-45). New York, NY: Oxfotd University Press, Inc.

Barbosa, P. A. (1996). At least two macrohymic units are necessary for modeling Brazilian Portuguese duration. *Proceedings of the 1st ESCA Tutorial and Research Workshop on Speech Production Modeling.* 4th Speech Production Seminar.

Barbosa, P., Duarte, M., & Kato, M. (2005). Null Subjects in European and Brazilian Portuguese. *Journal of Portuguese Linguistics*.

Bednarek, M., & McCarthy, M. (. (2011). The Routledge Handbook of Corpus Linguistics. *Australian Review of Applied Linguistics, 34* (1), pp. 104-107.

Biber, D., Johansson, S., Leech, H., Conrad, S., & Finegan, E. (2002, November). Longman grammar of Spoken and Written English. *English Language and Linguistics*, pp. 379-416.

Buhring, D. (1995). *The 59th Street Bridge Accent: On the Meaning of Topic and Focus.* Ph. D. dissertation. Tubingen: University of Tubingen.

Chomsky, N. (1957). *Syntactic Structures.* Berlin: Mouton de Gruyter (formerly Mouton, The Hague).

Chomsky, N. (1981). *Lectures on Government and Binding: The Pisa Lectures.* Holland: Foris Publications. Reprint. 7th Edition. Berlin and New York: Mourton de Gruyter, 1993.

Cinque, G. (1992). *A Null theory of Phrase and Compound Stress.* LI, 24.

Cooper, W. E., & R., H. J. (1975). World Order. In R. E. Grossman, L. J. San, & T. J. Vance, *Papers from the parasession on functionalism* (pp. 63-111). Chicago: Chicago Linstuistic Society.

Costa, J. (1996). Positions for Subjects in European Portuguese. *Proceedings of the West Coast Conference on Formal Linguistics XV.*

Costa, J. (1996). Scrambling in European Portuguese. *Proceedings of SCIL 8.* MIT Working Papers in Linguistics.

Costa, J. (1996). *Word order and constraint interaction. Unpublished Manuscript.* HIL/Leiden University.

Costa, J. (2000). Word Order and Discourse-Configurationality in European Portuguese. In J. Costa, *Portuguese Syntax: New Comparative Studies* (pp. 94-115). New York, NY: Oxford University Press.

Crystal, D. (1997). *The Cambridge Encyclopedia of Language, Second Edition.* Ernst Dlett Sprachen.

Cyrino, S. M. (1993). Observações sobre a mudança diacrônica no português do Brasil: objeto nulo e clíticos. In I. Roberts, & M. A. Kato, *Português Brasileiro: uma viagem diacrônica* (pp. 163-184). Campinas: Ed. da UNICAMP.

Cyrino, S. M. (1994). *O objeto nulo no Português do Brasil: um estudo sintático-diacrônico.* Ph.D Dissertation. Campinas: UNICAMP.

Davies, M. (1990-2011). *Corpus of Contemporary American English.* Provo, UT: Brigham Young University.

Davies, M. (2008, March 1). New Directions in Spanish and Portuguese Corpus Linguistics. *Studies in Hispanic and Lusophone Linguistics*, pp. 149-186.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 159-190.

Davies, M. (2010). More than Just a Peephole: Unisng large and diverse online corpora. *International Journal of Corpus Linguistics*, 412-420.

Davies, M. (2013, May 6). *corpus.byu.edu*. Retrieved from corpus.byu.edu: corpus.byu.edu

Davies, M., & Ferreira, M. (2006). *Corpus do Português.* Provo: Brigham Young University.

Downing, P., & Noonan, M. (. (1995). *Word Order in Discourse.* Philadelphia: John Benjamins Publishing Company.

Duarte, M. (1993). Do pronome nulo ao pronome pleno: a trajetória do sujeito no português do Brasil. In I. a. Roberts, *Português Brasileiro: Uma viagem diacrônivs (Homenagem a Fernando Tarallo)* (pp. 107-128). Campinas: Edotora da UNICAMP.

Duarte, M. (1995). *A Perda do Princípio "Evite pronome" no Português Brasileiro.* Ph.D. Dissertation. Campinas: UNICAMP.

Duarte, M. E. (1989). Clítico acusativo, pronome lexical e categoria vazia no português do Brasil. In F. (. Tarallo, *Fotografias Sociolinguísticas* (pp. 19-34). Campinas: Pontes.

Duarte, M. E. (n.d.). O sujeito pronominal no português coloquial europeu. In G. M. Silva, & S. Bortoni, *Fotografias Sociolinguísticas II.* Campinas: Pontes.

Estatística, I. B. (2013, April 30). *Censo Demográfico 2010*. Retrieved from IBGE: http://www.ibge.gov.br

Galves, C. C. (1987). A Sintaxe do Português Brasileiro. *Ensaios de Linguística, 13*, pp. 31-50.

Galves, C. C. (1991). *Agreement and Subjects in Brazilian Portuguese. ms.* Campinas: UNICAMP.

Group, T. W. (2013, May 8). *Data: Brazil*. Retrieved from The World Bank: http://data.worldbank.org/country/brazil

Haspelmath, M. (2001). The European linguistic area: Standard Average European. In M. Haspelmath, & e. al., *Language Typology and Language Universals, vol. 2* (pp. 1492-1510). Berlin and New York: Walter de Gruyter.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar.* MIT Press.

Kato, M. A. (1993). Recontando a história das relativas em uma perspectiva paramétrica. In I. Roberts, & M. A. Kato, *Português Brasileiro: uma viagem diacrônica* (pp. 223-306). Campinas: UNICAMP.

Kato, M. A. (1994). A ordem dos constituintes e os elementos portadores de traços-phi. Relatório de pós-doutorado. *Seminário da Gramática do Português Falado.* Campos do Jordão, SP: UCLA / FAPESP.

Kato, M. A., & Raposo, E. (1996). European and Brazilian Portuguese Word Order: Questions, Focus and Topic Constructions. In C. Parodi, C. Quicoli, & M. a. Saltarelli, *Aspects of*

Romance Linguistics: Selected Papers from the LSRL XXVI. (pp. 267-278). Washington: Georgetown University Press.

Kiss, K. É. (1998). Identification focus versus information focus. *Language 74(2)*, 245-273.

Larousse. (2008). *Portuguese Dictionary.* Boston: Houghton Mifflin.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (2013). Ethnologue: Languages of the World, Seventeenth edition. Dallas, Texas, SIL International. Online version: http://www.ethnologue.com.

Longman. (1999). *Longman Grammar of Spoken and Written English.* Cambridge: Cambridge University Press.

Martins, A. M. (1994). *Os Clíticos na História do Português.* Lisbon: Ph.D dissertation. University of Lisbon.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics: An introduction (2nd edition).* Edinburgh: Edinburgh University Press.

Nagata, H. (January, 1988). The Relativity of Linguistic Intuition: The Effect of Repetition on Grammaticality Judgments. *Journal of Psycholinguistic Research 17.1*, pp. 1-17.

Nash, L. (1995). *Argument Scope and Case Marking in SOV and in Ergative Languages: the case of Georgian.* Paris 8: Ph.D dissertation.

Pagotto, E. G. (1992). *A Posição dos Clíticos em Português: um estudo diacrônico.* Dissertação de mestrado. Campinas: UNICAMP.

Pagotto, E. G. (1993). Clíticos, mudança e seleção natural. In I. Roberts, & M. A. Kato, *Português Brasileiro: uma viagem diacrônica* (pp. 183-203). Campinas: UNICAMP.

Reinhart, T. (1995). *Interface Strategies.* Ms. OTS/Utrecht University.

Rizzi, L. (1982). *Issues in Italian Syntax.* Foris: Dordrecht.

Rooth, M. (1985). Association with focus. PhD Dissertation. Amherhest, MA: University of Massachusetts.

Santos, D. (1991). *Contrastive tense and aspect data.* Lisboa.

Silva, G. V. (2001). *Word Order in Brazilian Portuguese.* New York: Mouton de Gruyter.

Software, G. P. (2013 йил 26-April). From Graph Pad: graphpad.com/quickcalcs/chisquared2/

Takaie, H. (2002). A Trap in Corpus Linguistics: The Gap between Corpus-Based Analysis and Intuition-Based Analysis. In T. N. Saito, *English Corpus Linguistics in Japan* (pp. 111-130). Amsterdam: Rodopi.

Trujillo, F. (n.d.). Discourse Analysis and Grammar. In *Uses of Spoken and Written English* (pp. 1-2). University of Massachusetts.

Weisstein, E. W. (2013, May 14). *Chi-Squared Test*. Retrieved from A Wolfram Web Resource: http://mathworld.wolfram.com/Chi-SquaredTest.html

Welkowitz, J., Cohen, B. H., & Ewen, R. I. (2006). *Introductory Statistics for the Behavioral Scientist.* Wiley.

APPENDIX

Null Subject Examples from the CDP:

(1) a. Academic

S V O

Ø ***trav-ou*** ***relações intelectuais.***

NS lock-3SG.PST relations intellectual

‘(He) locked intellectual relations...’

b. Oral

S V O

Ø ***esij-o*** ***respeito em relação aos horários.***

NS demand-1SG.PRS respect in relation to.the hours

‘(I) demand respect with regards to the schedule.’

c. Fiction

S V O

Ø ***assist-i*** ***a ocupação alemã.***

NS see-1SG.PST the occupation German

‘(I) saw the German occupation.’

c. Newspaper

S V O

Ø ***mostr-a*** ***a vida de um policial.***

NS show-3SG.PRS the life of a policeman

‘(It) shows the life of a policeman.’

Covert Object Examples from the CDP

(2) a. Academic

S V O

Toda estratificação implic-a **Ø**

every stratification implicate-3SG.PRS CO

‘Every stratification implies (it).’

b. Oral

S V O

Eu abr-o **Ø** ***em Dezembro.***

I open-1SG.PRS CO in December

‘I open (it) in December.’

c. Fiction
S V O
Todo mundo confer-iu Ø
All world confirm-3SG.PST CO
'Everyone confirmed (it).'

c. Newspaper
O V S
Ø vai valer a determinação de cada
CO go.3SG.PRS to.be.worth the determination of each
'Everyone's determination will be worth (it).'

Word-Order Variation in the CDP

(3) a. Fiction
S V O
o navio passava uma série de canaviais verde-claros
the ship pass-3SG.PST a series of reeds green-light
'The ship passed a series of light-green reeds'

b. Newspaper
V O S
ganh-a importância neste cenário o leilão
gain-3SG.PRS importance in.this scenario the auction
'The auction gains importance in this scenario.'

c. Fiction
O V S
outro tanto de elogios$_i$ teve o serviço de coquetel.
other as.much of praise had-3SG.PST the service of cocktail
'Just as much praise$_i$ received the cocktail service' (lit.).