UNIVERSITY OF WOLVERHAMPTON
School of Law, Social Sciences and Communications
UNIVERSIDADE DO ALGARVE
Faculdade de Ciências Humanas e Sociais

Simone Pereira

Linguistics Parameters for Zero Anaphora Resolution

Project submitted as part of the programme of study for the award of

MA in Natural Language Processing
& Human Language Technology

Supervisors:

Jorge Baptista
Richard Evans

May 2010

Linguistics Parameter for Zero Anaphora Resolution

Simone Pereira

Supervisors: Jorge Baptista and Richard Evans

Project submitted as part of the programme of study for the award of MA in Natural Language Processing & Human Language Technology

“Revised version after the presentation”

Wolverhampton

May 2010

UNIVERSITY OF WOLVERHAMPTON
SCHOOL OF LAW, SOCIAL SCIENCES AND COMMUNICATIONS
MA NATURAL LANGUAGE PROCESSING & HUMAN LANGUAGE TECHNOLOGY

Name: SIMONE CRISTINA PEREIRA

Date: 26/05/2010

Title: LINGUISTICS PARAMETERS FOR ZERO ANAPHORA RESOLUTION

Module Code: LN4007

Presented in partial fulfillment of the assessment requirements for the above award

Supervisors: JORGE BAPTISTA
RICHARD EVANS

Declaration:

This work or any part thereof has not previously been presented in any form to the University or to any other institutional body whether for assessment or for other purposes. Save for any express acknowledgements, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

This project did not involve contact with human subjects, and hence did not require approval from the LSSC Ethics Committee.

Signed:_____ Date:_____

To my family, supervisors
and all people who contributed
for the realization of this study

Abstract

This dissertation describes and proposes a set of linguistically motivated rules for zero anaphora resolution in the context of a natural language processing chain developed for Portuguese. Some languages, like Portuguese, allow noun phrase (NP) deletion (or zeroing) in several syntactic contexts in order to avoid the redundancy that would result from repetition of previously mentioned words. The co-reference relation between the zeroed element and its antecedent (or previous mention) in the discourse is here called zero anaphora (Mitkov, 2002). In Computational Linguistics, zero anaphora resolution may be viewed as a subtask of anaphora resolution and has an essential role in various Natural Language Processing applications such as information extraction, automatic abstracting, dialog systems, machine translation and question answering. The main goal of this dissertation is to describe the grammatical rules imposing subject NP deletion and referential constraints in the Brazilian Portuguese, in order to allow a correct identification of the antecedent of the deleted subject NP. Some of these rules were then formalized into the Xerox Incremental Parser or XIP (Ait-Mokhtar et al., 2002: 121-144) in order to constitute a module of the Portuguese grammar (Mamede et al. 2010) developed at Spoken Language Laboratory (L2F). Using this rule-based approach we expected to improve the performance of the Portuguese grammar namely by producing better dependency structures with (reconstructed) zeroed NPs for the syntactic-semantic interface. Because of the complexity of the task, the scope of this dissertation had to be limited: (a) subject NP deletion; b) within sentence boundaries and (c) with an explicit antecedent; besides, (d) rules were formalized based solely on the results of the shallow parser (or chunks), that is, with minimal syntactic (and no semantic) knowledge. A corpus of different text genres was manually annotated for zero anaphors and other zero-shaped, usually indefinite, subjects. The rule-based approached is evaluated and results are presented and discussed.

Keywords: Anaphora resolution, zero anaphora, linguistically-motivated rule-base approach, Brazilian Portuguese.

Resumo

Este estudo descreve e apresenta um conjunto de regras linguisticamente motivadas para a resolução de anáfora zero no contexto de uma cadeia de processamento de linguagem natural desenvolvida para o Português. Certas línguas, como o Português, permitem o apagamento (ou redução a zero) de grupos nominais (GN) em vários contextos sintácticos a fim de evitar a redundância que resultaria da repetição de elementos previamente mencionados no discurso. A relação de correferência entre o elemento reduzido a zero e o seu antecedente (ou menção anterior) no discurso é aqui chamada de anáfora zero (Mitkov 2002). Em Linguística Computacional, a resolução de anáfora zero pode ser vista como uma subtarefa da resolução de anáfora em geral, e tem um papel essencial em várias aplicações em Processamento de Linguagem Natural, tais como extracção de informação, sumarização automática, sistemas de diálogo, tradução automática ou resposta automática a perguntas. O principal objectivo deste estudo consiste na descrição das condições gramaticais que impõem a redução a zero de grupos nominais e as respectivas restrições de correferência, no Português do Brasil, de forma a permitir uma correcta identificação do antecedente de sujeitos reduzidos a zero. Algumas destas regras foram então formalizadas de modo a constituírem um módulo de resolução anáfora integrado na gramática do Português (Mamede et al., 2010) desenvolvida no Spoken language Laboratory (L2F) para Xerox Incremental Parser, ou XIP (Ait-Mokhtar et al., 2002: 121-144). Utilizando esta abordagem baseada em regras, pretende-se melhorar as estruturas de dependências extraídas das frases reconstituindo e representando os GN reduzidos a zero. Devido à complexidade da tarefa, este estudo limitar-se-á: (a) a GN sujeitos reduzidos a zero; (b) no âmbito intrafrásico; e (c) com um antecedente explícito; além disso, as regras de reconstituição de sujeito basear-se-ão exclusivamente nos resultados da cadeia de processamento, em particular numa análise sintáctica superficial (chunking), ou seja, com um mínimo de conhecimento sintáctico (e sem conhecimento semântico). Um corpus de diferentes géneros textuais foi manualmente anotado de forma a identificar as situações de anáfora zero bem como outros tipos de sujeito elíptico, geralmente indefinidos. Esta abordagem baseada em regras foi avaliada e os resultados são apresentados e discutidos.

Palavras-chave: Resolução de anáfora, anáfora zero, abordagem baseada em regras linguisticamente motivadas, Português do Brasil.

This project was supported by the European Commission, Education & Training, Erasmus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT programme.

Table of Contents

Symbols and abbreviations

BP	Brazilian Portuguese
EP	European Portuguese
	Differences among the two varieties of Portuguese are signalled by raised $^{ep}/_{bp}$: $^{ep/bp}$*shopping* / $^{ep/*bp}$*centro comercial*
NLP	Natural Language Processing
AR	Anaphora Resolution
ZAR	Zero Anaphora Resolution
ZA	Zero Anaphora
XML	Extensible Mark-up Language
Ø	zeroed constituent
( )	optional constituent
+	separates elements between (...) that can appear in a given syntactic position
X_i	index of coreference: the constituent X_i is coreferent of another constituent Y_i
?	dubiously acceptable sentence
*	unacceptable sentence
“ ”	free translation
' '	word-for-word translation
N	noun
NP	noun phrase
PP	prepositional phrase
[he]	in the examples, signals zeroed elements reconstituted for clarity sake in the translation
SC	subordinate clause
CC	coordinate clause
MC	main clause
AC	adverbial subordinate clause
m	masculine
f	feminine
sg	singular
pl	plural
1,2,3	person (first, second, third)

List of Figures

List of Tables

1 Introduction

Some features characterize a set of words as a text. One of these features is cohesion. Cohesion occurs where the interpretation of some element in the discourse is dependent of another, i.e. an element cannot be decoded in the text except by recourse to another element that it presupposes (Halliday and Hasan, 1976: 4). For example:

(1.1) *Wash and core* six cooking apples$_i$. *Put* them$_i$ *into a fireproof dish* (Halliday and Hansan, 1976: 2)

The element *them* (in the second sentence) presupposes for its interpretation the element *six cooking apples* (in the first sentence). When the presupposing and the presupposed element are resolved the cohesion between the two sentences is established.

There are different types of cohesion and our interest, in this dissertation, is with a particular type of cohesion mechanism. In some linguistic situations, repeated mentions of NPs, usually already present in a previous utterance or in a previous constituent of the same utterance may be reduced to pronoun or to zero (NP deletion) in order to avoid redundancy from repetition (Harris, 1991: 6).

(1.2) *John *went to school and then* John *went to the mall*

(1.3) John *went to school and then* (he went) *to the mall*

In sentence (1.2) the word *John* cannot occur in the second clause because it is not recommended that the same entity be mentioned twice within the same sentence. This recommendation is made through the rule called pronominalization which governs the process of reference. It is the sentence structure which determines, within limits, when the second mention of the entity will be named again or it will be referred to by a pronoun (Halliday and Hasan, 1976: 8).

In sentence (1.3) the words *he went* (in the second clause) may or not occur. The writer chose not to use the pronoun and the verb in order to avoid redundancy. She or he may also keep the verb while zeroing the pronoun (1.4), but not the opposite (1.5).

(1.4) *John went to school and went to the mall*

(1.5) **John went to school and he to the mall*

The reduction of the repeated NPs to zero is our object of study on this dissertation.

1.1.1 The terminology adopted

According to Harris (1991: 5) all instances (or discourses) of a language are word sequences which satisfy certain combinatory constraints. One sentence would be a reconstruction of their unreduced form. Certain sentences contain in a regular way the same component (words-sequence) as other sentences which are paraphrastic to them. There are two cases of these paraphrases: a) many sentences consisted simply of other sentences plus additional words, with the meaning of the included sentence being both preserved and added to, and b) many sentences consist of another sentence with no additional elements but with a change, in most cases a reduction or a transformation, that leaves the meaning of the source sentence unaltered. For each language there is a particular set of reductions and particular conditions necessary for their being carried out (Harris, 1991: 7).

For example, the coordination of constituents is explained through a general rule which determines that two or more *quasi* identical sentences should be merged except for the constituents that have to be coordinated, creating, thus, a sentence with the coordinated constituent in it (Harris, 1975: 174). Therefore according to this point of view, sentence (1.3) derives from two base sentences:

(1.6) *John went to the school*

(1.7) *He (John) went to the mall*

The result of the reduction after the coordination of these sentences is the sentence (1.3) without the words *he went*. Thus, sentence (1.3) is only the reduction of a longer, unreduced sentence.

Chomsky (1981) already defined this kind of phenomenon as a characteristic of languages in which certain classes of pronoun may be omitted when they are in some sense pragmatically inferable. This kind of languages is called by him as *pro-drop* (pronoun dropping) languages.

Several languages are considered pro-drop. Among them languages such as Japanese allow the pronoun deletion not for only subject but for practically any structural position. Romance Languages such as Spanish, Italian, and Portuguese are considered partially pro-drop because they allow pronoun deletion in several, syntactic constrained contexts.

Languages like English and French are considered as non-pro-drop languages because in most of the cases the pronoun deletion is not allowed. However in a few

cases the pronoun can be dropped, as for example, in imperatives sentences (when someone gives an order) and in informal speech.

Halliday and Hasan (1976) designate this kind of cohesion mechanism as *ellipsis*. According to them ellipsis is the omission of an item. This phenomenon has a relation within the text, and in the great majority of instances the presupposed item is present in the preceding text (Halliday and Hasan, 1976: 144). So the omission of the words in sentence (1.3) is classified as ellipsis by the authors.

Finally, Mitkov (2002) name this phenomenon of the omission of a word as *zero anaphora* or ellipsis. Accordingly zero anaphors are 'invisible' anaphors, i.e. the anaphors do not appear to be in the sentence because they are not overtly represented by a word or phrase. Since one of the properties and advantages of anaphora is its ability to maintain the amount of information presented via an abbreviated linguistic form, ellipsis may be "the most sophisticated variety of anaphora" (Mitkov, 2002: 12).

As ellipsis is associated with the deletion of linguistic forms, the correct coherence of a sentence or a discourse segment imposes the recovery of the meaning via its antecedent. Thus, the phenomenon presented in sentence (1.3) is called zero anaphora.

On this dissertation we adopted the same terminology used by Mitkov.

1.1.2 The language studied

The language studied on this dissertation is the Brazilian Portuguese. Portuguese, in general, has a very rich verbal inflection[1], and the deleted subject can easily be recovered through verbal inflection.

The grammatical rules governing NP deletion may vary among languages, and even among different varieties of the 'same' language, as in the case of Brazilian (BP) vs. European Portuguese (EP):

[1] In Portuguese verbs have a very rich inflectional morphology. Usually, verbs distinguish almost every person-number variations. The subject may often be zeroed since it can easily be reconstructed from the verb ending:

(*Eu*) compr*o*	I buy_1sg
(*Tu*) compr*as*	You buy_2sg
(*Ele*) compr*a*	He buys_3sg
(*Nós*) compr*amos*	We buy_1pl
(*Vós*) compr*ais*	You buy_2pl
(*Eles*) compr*am*	They buy_3pl

Besides the 6 person-number pronouns described above, Portuguese has, also, another form, *você* (you – singular) and *vocês* (you – plural). These pronouns refer to the addressee (2sg/ 2pl) but impose 3sg/3pl on verbal agreement. This equivalence is systematic.

(1.8) **O João*$_i$ *foi à escola e depois o João*$_i$ *foi ao centro comercial/shopping*

*'John$_i$ went to school and then John$_i$ went to the mall'

(1.9) *O João*$_i$ *foi à escola e depois (Ø*$_i$ + $^{*ep,\ pb}$*ele*$_i$*) foi ao centro comercial/shopping*

'John$_i$ went to school and then (Ø$_i$ + $^{*ep,\ pb}$ he$_i$) went to the mall'

(1.10) *O João*$_i$ *foi*$_j$ *à escola e depois Ø*$_{ij}$ *ao centro comercial/shopping*

'John went to school and then to the mall'

In sentence (1.8) the NP *O João* 'John' cannot occur in the second clause because the same entity was already referred in the first clause (pronominalization rule). In sentence (1.9) the pronoun *ele* 'he' can be zeroed (marked with the symbol Ø) both in European Portuguese and in Brazilian Portuguese but the pronoun can occur only in Brazilian Portuguese; in sentence (1.10) the reduction of the verb *foi* 'went' imposes the subject NP deletion (ele 'he'). Hence in Brazilian Portuguese, both to pronoun and to zero can occur, while in European Portuguese only zero-reduction is allowed.

The term *anaphor* is used to designate the pronoun in NP reduction or the syntactic slot left empty by NP deletion; in the case of the sentence (1.10) the term anaphor is marked by the symbol Ø. On the other hand, the term *anaphora* is a general term for the referential relation between the anaphor and its antecedent. It includes both anaphora proper: (i) when the antecedent appears in a previous moment in the discourse, e.g. in sentence (1.11) the NP *João e Maria* 'John and Mary' appears before the symbol Ø; and (ii) when the antecedent appears in a later moment in the discourse (called *cataphora*), e.g. in the sentence (1.12), the symbol Ø appears before the NP *o óvulo* 'the ovum'.

(1.11) *João e a Maria*$_i$ *viajaram para o Sul mas Ø*$_i$ *não foram de férias*

'John and Mary travelled to the South, but [they] were not in vacation'

(1.12) *Caso Ø*$_i$ *não seja fecundado*, *o óvulo*$_i$ *morrerá*

'If [the ovum] is not fertilized, the ovum will die'

Subject NPs in Portuguese are traditionally classified[2] into the following types:

a) explicit subject

Explicit subjects NPs include simple and coordinated NPs depending on the head being a single N or coordinated NPs. Naturally, in this type, the subject NP is explicit and the zeroed subject NP does not occur.

[2] For an overview of Portuguese grammar on subject types, please refer to Cunha and Cintra (1984: 125-133); Bechara (2001: 408-414); Brito and Matos (2003: 435-449); among others.

(1.13) A Maria *comprou um livro* ‘Mary_*3sg* buys a book’

(1.14) O João e a Maria *compraram um livro* ‘John and Mary_*3pl* buy a book’

b) indefinite subject

Following Cunha, C. and Cintra, L. (1984: 128-129) two types of indefinite subject are considered:

i. verb in third singular person:

- indefinite clitic pronoun *–se*

This pronoun is equivalent to an indefinite subject NP such as indefinite pronoun *alguém* ‘someone’.

The clitic imposes a 3rd person singular agreement to the main verb.

(1.15) *Precisa-se de empregados* ‘(One / Someone) needs employees’

≈ *(Alguém) precisa de empregados*

- passive particle

This is however a passive-like construction where an object NP is raised to the subject position, the (transitive) verb agrees with the new subject NP and, usually, the former subject is omitted.

(1.16) *Compraram-se vários livros* ‘(Someone) bought_*3rdpl* several books_*pl’*

In spite of post verbal position, only the plural NP can account for verbal agreement. Traditionally the *–se* form is called a passive particle.

Naturally when the subject NP is in the singular, an ambiguous sentence is produced:

(1.17) *Comprou-se um livro*

‘(Someone) bought_*3rdsg* a book_*sg’* / “A book was bought”

In this case the *–se* form can be analysed both as an indefinite clitic pronoun and as the passive particle.

Anyway either the clitic pronoun or the passive-like sentences have an explicit formal subject, therefore these cases fall out of the scope of this dissertation.

ii. verb in third person plural:

(1.18) *Deixaram um presente na minha mesa*

‘Someone left_*3rdpl* a gift on my desk’

In sentence (1.18) the action was made by someone, but the subject cannot be recovered because it is an indefinite subject. Anaphora resolution in this case should be blocked.

This subject type cannot be resolved by purely syntactic analysis. Since such

world-knowledge, *e.g.* pragmatics and linguistics information is required. Moreover as this sentence type is used mainly in oral language and in colloquial register, it will not be deal with here.

c) impersonal subject

Lexically determined verb constructions deal with in this section is traditionally classified as “impersonal”. These constructions concern:

i. existential constructions with *haver* ‘there is’:

(1.19) *Há muitos livros na biblioteca* ‘There are many books in the library’

ii. meteorological phenomena:

This kind of verbs denotes some nature phenomenon like: *chover*, *nevar* ‘to rain, to snow’ (1.20) or the corresponding verb-noun constructions (1.21).

(1.20) *Nevou ontem a noite* ‘It snowed last night’

(1.21) *No sul do Brasil, faz (noites muito frias + nevoeiros + sol) no inverno*

‘In the south of Brazil, it makes (very cold nights + fog + sun) in winter’

iii. part-of-day expressions:

(1.22) *(Amanheceu + entardeceu + anoiteceu) tarde*

‘It (dawned + grew dark) later’

iv. formulaic expressions concerning time, hours and dates:

Formulaic expressions of time with verb *fazer* ‘to make’ (1.23) and verb *haver* ‘there be’ (1.25):

(1.23) *Ontem, (fez + fizeram$^{bp/*ep}$) dez anos que ele morreu*

‘Yesterday, it was done ten years since he died’

“It is ten years yesterday since he died”

(1.24) Ele morreu (faz + *fizeram) dez anos ‘It is ten years since he died’

In Brazilian Portuguese it is acceptable the verb in the plural forms with the construction verb + NP + that + phrase. But in the construction verb + NP the verb in the plural form it is not acceptable.

(1.25) *Há quinze dias, Maria esteve em São Paulo*

‘Two weeks ago, Mary was in São Paulo’

Duration expressions with verbs *ser* or *estar* ‘to be’ (1.26):

(1.26) O tempo de espera para uma consulta (são + é de) dois meses

‘The time of waiting for an appointment it is two month’

Formulaic expressions of hours with verb *ser* or *estar* ‘to be’:

(1.27) *(É + São) duas horas da tarde* ‘It is two o’clock in the afternoon’

Formulaic expression indicating dates with verb *ser* 'to be' (1.28) and verb *estar* 'to be' (1.29):

(1.28) *É primeiro de setembro* 'It is September first'

(1.29) *Estamos (^{bp,ep}em + ^{ep}a) 3 de Abril* (Ranchhod, 1990:77)

'We are in April 3' "It is April 3"

Naturally, impersonal constructions do not concern anaphora resolution since there is no coreference involved. Nevertheless they must be signalled during text processing.

d) non-explicit, hidden subject

In Portuguese, pronominal 1st and 2nd person subject NPs are usually reduced to zero.[3]

(1.30) *(Eu) Comprei um livro* 'I_*1sg* bought a book'

(1.31) *(Tu) Compraste um livro* 'You_*2sg* bought a book'

(1.32) *(Nós) Compramos um livro* 'We_*1pl* bought a book'

(1.33) *(Vós) Comprastes um livro* 'You_*2pl* bought a book'

3rd person pronominal subjects cannot be reduced unless coreference with previous instance of the same entity can be recovered:

(1.34) *João e Maria$_i$ foram ao shopping e Ø$_i$ compraram um livro*

'John and Mary went to the mall and they bought a book'

In sentence (1.34), the zeroed subject of the second verb *compraram*, 'bought_3pl' can be recovered[4] from its previous occurrence in the (same) utterance.

In this dissertation only deleted, 'non-explicit' or 'hidden' subjects will be considered.

1.2 Motivation

Anaphora Resolution (AR) has an important role in understanding the information that is embedded but it is not explicit in the discourse. This study aims to facilitate machine understanding of the information conveyed by natural language.

In Computational Linguistics, AR has an essential role in various NLP applications such as information extraction, automatic abstracting, dialogue system, machine translation and question answering (Mitkov, 2003: 275). For an example,

[3] Several pragmatic conditions govern different meaning (focus) associated with this variations.

[4] Traditional Portuguese grammars such as Bechara (2001) and Mateus *et al* (2003) do not consider hidden or non-explicit subject as ellipsis. We will not discuss their theoretical point of view here.

machine translation systems need to understand the discourse information to perform adequate translations. This did not happen with the majority system developed in the 1970s and 1980s (Mitkov, 2003: 275-276):

> "Unfortunately, the majority of MT system develop in the 1970s and 1980s did not adequately address the problems of identifying the antecedents of anaphors in the source language and producing the anaphoric 'equivalent' in the target language. As a consequence, only a limited number of MT systems have been successful in translating discourse, rather than isolated sentences. One reason for this situation is that in addition to anaphora resolution itself being a very complicated task, translation adds a further dimension to the problem in that the reference to a discourse entity encoded by a source language anaphor by the speaker (or writer) has not only to be identified by the hearer (translator or translation system), but also re-encoded in a different language." (Mitkov, 2003: 275-276)

Zero anaphora resolution (ZAR) may be viewed as a subtask of AR. In languages like Portuguese, Spanish, Italian, Polish, Chinese, Japanese, Korean or Thai (Mitkov, 2002: 13) zeroed NP subjects are widely used and this requires the adequate resolution of zero anaphors, which is not simple.

For example, some constructions present problems in the recovering of the zeroed NP subject:

(1.35) *A Maria*$_{i}$_f_sg *disse*_sg *à amiga*$_{j}$_f_sg *que* Ø$_{i/j}$ *estava*_3sg *apaixonada*_f_sg

'Mary told her friend that was in love'

(1.36) *A Maria e o João*$_{i}$_m_pl *disseram*_pl *aos amigos*$_{j}$_m_3pl *que* Ø$_{i/j}$ *estavam*_3pl *apaixonados*_m_pl

'Mary and John told their friends that were in love'

Sentences (1.35) and (1.36) present an ambiguous situation. The conjunction *que* 'that' can be an integrant conjunction[5] or a relative pronoun. If the integrant conjunction is considered, then the subject of the verb *estavam_3sg* 'be' will be *Maria* 'Mary' in (1.35) and the coordinated NPs *A Maria e o João* 'Mary and John' in (1.36). But, if the word *que* 'that' is considered relative pronoun, then the subject will be the indirect (dative) object *à amiga* 'friend' in (1.35) and *aos amigos* 'friends' in (1.36).

If valence information is available for the main verb *dizer* 'say' it might be possible to parse the subclause correctly and hence solve the zero anaphora adequately. In this case, a preferential analysis results from the fulfilling of all syntactic slots of the main verb (e.g. *dizer* 'to say'), since with the relative, the absence of direct object renders the sentence unacceptable.

[5] For a definition of integrant conjunction, please refer to chapter 3.

Gender-number agreement can also be useful to solve anaphora:

(1.37) *A Maria*$_i$_f_sg *disse*_sg *ao amigo*_m_sg *que* Ø$_i$ *estava*_sg *apaixonada*_f_sg

'Mary told her friend that was in love'

(1.38) *A Maria e o João*$_i$_m_pl *disseram*_pl *ao amigo*_m_sg *que* Ø$_i$ *estavam*_pl *apaixonados*_m_pl

'Mary and John told their friend that were in love'

(1.39) *A Maria e o João*$_i$_m_pl *disseram*_pl *às amigas*_f_pl *que* Ø$_i$ *estavam*_pl *apaixonados*_m_pl

'Mary and John told their friends that were in love'

In sentences (1.37), (1.38) and (1.39), the conjunction *que* 'that' is classified as an integrant conjunction and, through the gender-number agreement, the adjective *apaixonada_f_sg* 'in love' or *apaixonados_m_pl* 'in love' indicates that the zeroed NP subject is the subject of the first clause – *A Maria* 'Mary' in sentence (1.37), *A Maria e o João* 'Mary and John' in sentences (1.38) and (1.39).

Consequently, in sentences in which the subject is zeroed, it is necessary to recover this subject because the information presented in the clause can be different according to the subject. This is not an easy task because the recovery of the zeroed NP subject involves different syntactic knowledge. The ZAR has an important role in languages that have zeroed NP subjects.

1.3 Goal

The main goal of this dissertation is to describe the grammatical rules imposing subject NP deletion in Brazilian Portuguese and its formalization so that a parser, using those rules, may correctly identify the antecedent of the deleted NP.

Identification of the antecedent of a deleted subject NP (zero anaphor) can be viewed as a module of the anaphora resolution task (Mitkov, 2002). Using this rule-based approach, we expect to improve the general performance of the Portuguese grammar (Mamede et al., 2010) developed for Xerox Incremental Parsing (XIP) (Ait-Mokhtar et al., 2002: 121-144) at L2F[6] in the INESC_ID[7] Lisbon, namely by producing better dependency structures with reconstructed zeroed NPs for the

[6] Spoken Language Laboratory: https://www.l2f.inesc-id.pt/wiki/index.php/Main_Page

[7] Institute for System and Computer Engineering Research and Development in Lisbon: http://www.inesc-id.pt/

syntactic-semantic interface.

The XIP parser is a formalism that integrates a number of description mechanisms for shallow and deep robust parsing, ranging from part-of-speech disambiguation, named entity recognition and chunking to dependency grammars. The system parses a text in the following steps: a) a pre-processing step, which includes text segmentation (tokenization and sentence splitting) and morphological analyses; b) a disambiguation step where words with more than one morphological category are disambiguated; c) a shallow parsing step (chunking); and d) a deep parsing stage where the dependencies among chunks and constituents are extracted.

The parse tree presents for each word the disambiguated morphological category, like, for example, the category ART for articles, NOUN for common or proper noun and so on. In the shallow parser, words are grouped in chunks like NP for noun phrase, ADVP for adverbial phrase, etc. In the deep parser, the system, based on linguistic rules, extracts dependencies among chunks. Dependency relationships can connect nodes according to specific relationships, typically standard syntactic dependencies, but also broader relationships, including relationships across sentences (Mamede et al, 2010: 4). For an example, the dependency DETD links a nominal head and a determiner, the dependency PREPD links the head of the PP to the preposition and so on.

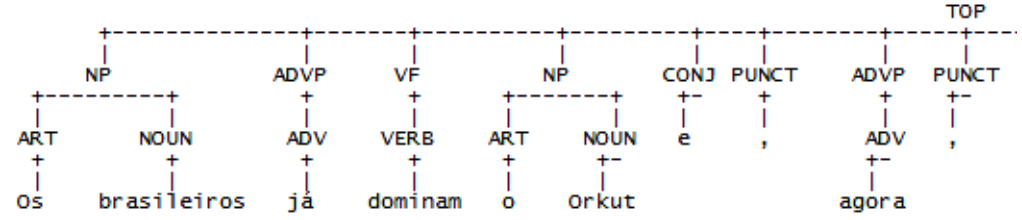
Consider the follow sentence:

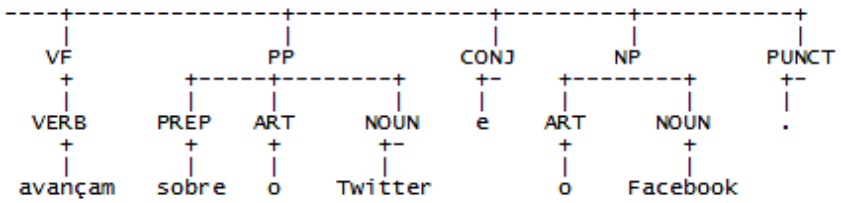
(1.40) *Os brasileiros*$_i$ *já dominam o Orkut e, agora,* Ø$_i$ *avançam sobre o Twitter e o Facebook*

'The Brazilians have mastered the Orkut and now [they] are encroaching on Twitter and Facebook'

The parse tree produced by the parser is presented in Figure 1:

Figure 1: Parse tree for the sentence 1.40





where the following dependencies were extracted (Figure 2):

Figure 2: Dependencies extracted for the sentence 1.40

MAIN(dominam)
HEAD(brasileiros,Os brasileiros)
HEAD(Orkut,o Orkut)
HEAD(Facebook,o Facebook)
HEAD(Twitter,sobre o Twitter)
HEAD(dominam,dominam)
HEAD(avançam,avançam)
HEAD(já,já)
HEAD(agora,agora)
HEAD(e,e)
HEAD(e,e)
DETD(brasileiros,Os)
DETD(Orkut,o)
DETD(Twitter,o)
DETD(Facebook,o)
PREPD(Twitter,sobre)
PREPD(Facebook,sobre)
COORD(e,Twitter)
COORD(e,Facebook)
VDOMAIN(dominam,dominam)
VDOMAIN(avançam,avançam)
MOD_PRE(dominam,já)
MOD_PRE(avançam,agora)
MOD_POST(avançam,Twitter)
MOD_POST(avançam,Facebook)
SUBJ_PRE(dominam,brasileiros)
CDIR_POST(dominam,Orkut)

As we can see, the parser found only the subject of the verb *dominam* 'have mastered' (in the dependencies appear the category SUBJ_PRE(dominam, *brasileiros*)). The zeroed NP subject of the verb *avançam* 'encroach' was not captured.

1.3.1 Presentation of XIP

XIP grammars have been developed for a number of languages, including French and English. The French grammar has been evaluated by Ait-Mokhtar et al (2002: 139)[8]. The parser is used in a several applications including an anaphora resolution system obtaining as result 74.8% (success rate) (Trouilleux, 2002).

For Portuguese the grammar has been developed under the collaboration between L2F laboratory at INESC_ID Lisbon and XRCE[9]. The Portuguese grammar has been used in a number of applications, and some of them were internationally evaluated. In Mendes et al. (2007) the XIP was integrated in a question-answering

[8] The evaluation of the French grammar was made using a corpus with 7.300 sentences of 23 words on average. In the evaluation of the linguistic performance, it was measured the precision and the recall of the subject dependency and direct object complement. For subject, precision and recall were respectively 93.45% and 89.36%, while the figures for verb complements were 90.62% and 86.56%.
[9] Xerox Research Centre Europe: http://www.xrce.xerox.com/

system – called QAL2F[10]. The system was evaluated at CLEF[11] (Peters, et al., 2007) having Portuguese as the query and target language. In Hagège et al. (2008) the XIP was integrated in named entity recognition system for Portuguese. The evaluation of this system was made during the second HAREM evaluation campaign (Mota and Santos, 2008). Comparing with other systems, results, especially the F-measure, reported by Hagège et al., can be considered as good; in general, the system had the third best result and, in the specific task for the recognition of named entities (NER) indicating time, the system presented the best performance in the contest. More recently, in Hagège et al. (2010), the module of NER for time expressions is revised and improved using XIP as the parser and NE extractor.

1.4 Structure of this document

This dissertation consists of 5 chapters and it is structured as follows:

- As the ZAR may be viewed as a subtask of AR, in the chapter 2 it is presented the literature review on AR and on ZAR.
- On chapter 3 it is presented the scope and the methods of this dissertation, the corpus developed for this study and the rules formalized in the XIP parser.
- The evaluation and the types of errors are presented on Chapter 4.
- The last Chapter, the Chapter 5, contains the conclusion and the future work.

[10] Question-answering system developed at L2F, INESC-ID.
[11] http://www.clef-campaign.org/

2 Related work

2.1 Grammars on subject NP deletion

This section will present how traditional grammars deal with zeroed subject. We chose to show the point of view of Portuguese and English grammars because the attention given to this phenomenon is quite different. While in Portuguese — a language where the zeroed NP subject is widely used – grammars explain only briefly this phenomenon without giving a comprehensive overview of the circumstances where it takes place; in English — a language that seldom features zeroed NP subject – traditional grammars present a more detailed study of this phenomenon.

2.1.1 Portuguese

Portuguese traditional grammars usually frame the study of subject NPs under the scope of agreement rules. When there is more than one clause per sentence and in the second clause there is a non-explicit subject, the verbal agreement is usually made according to the NP's head of the main clause. In sentence (2.1), the verb *voltaram* 'come back' agree with the subject *João e Maria* 'John and Mary'.

(2.1) *João e Maria*$_i$ *foram ao cinema e depois Ø*$_i$ *voltaram para casa*

'John and Mary went to the cinema and after they come back home'

Some grammars (Brito et al., 2003) deal with zeroed NP subjects under the topic of the sentence structure and sentence types. Again, Portuguese is presented as a null-subject language and this linguistics feature is explained by the rich verbal inflection.

Nevertheless, attention is drawn to the fact that in many verbs there is often a systematic 1sg/3sg homograph, which hinders the reference resolution procedure:

(2.2) *Cantava muito naquele Verão* (*idem*: 442)

'I sang a lot that summer or he sang a lot that summer'

In this sentence, without any previously clue, it is not possible to discover the zeroed NP antecedent.

Other grammars (Bechara, 2001), the zeroed NPs subjects are described in the chapter concerning the sentence and the sentence functions. The author deals with

the zero NPs subjects as an optional and non-optional term. He said that some terms can be zeroed or because this term was already used before, or because this term can be recovered easily through the context which the sentence is inserted.

Both Portuguese grammars, Bechara (2001) and Matos (2003) do not consider the non-explicit subjects as ellipses. They justify this statement saying the non-explicit subject can be recovered through the verbal inflectional.

> "A necessidade de explicitação do sujeito gramatical mediante um sujeito explícito é ditada pelo texto; a rigor, portanto, não se trata de "elipse" do sujeito, mas do "acréscimo" de expressão que identifique ou explicite a que se refere o sujeito gramatical indicado na desinência do verbo finito ou flexionado. Em português, salvo casos de ênfase ou contraste, não se explicita o sujeito gramatical mediante os pronomes de 1.ª e 2.ª pessoas do singular e do plural (...)"[12] (Bechara, 2001: 592)

In Matos 2003, the author presents the same idea:

> "(...) o constituinte não realizado, o Sujeito Nulo, seja interpretável independentemente de qualquer expressão linguística ou situacional prévia, bastando a presença das marcas de concordância verbal para ser recuperado (...)"[13] (Matos, 2003: 872)

2.1.2 English

In English, the subject is an obligatory element. Only in certain specific constructions like non-finite subordinated clauses (2.3) and imperatives (2.4) are subject NPs be omitted.

(2.3) *I expected him to go*

(2.4) *Leave your coat in the hall*

Some constructions use the pronoun it only to satisfy the syntactic need for a subject but has no identifiable meaning.

> "The fact that the subject is obligatory is reflected in the possibilities for reducing clauses when material is recoverable from the context. *Sue has eaten then already*, say, can be reduced to *She has* (e.g. in answer to the question *Has Sue eaten already*?), but not to **Has* or **Has eaten*. *She has* is what we will refer to as a **maximal finite reduction**, i.e. a finite clause that can't be reduced any further, and this construction must contain a subject together with an auxiliary or the pro-form do." (Huddleston and Pullum 2002: 239)

[12] "The need of explanation of the grammatical subject by an explicit subject is dictated by the text; so this phenomenon is not considered as subject 'ellipse', but it is considered as an addition of the words that identify or explicit the grammatical subject indicated by the finite verbal ending or by the verbal inflection. In Portuguese, except for the cases indicating emphasis or contrast, the grammatical subject for the pronoun of first and second singular or plural person is not explicit (…)" – free translation.

[13] "(…) the constituent unrealized, the null subject, is interpretable independently of any previous linguistic expression or situational, sufficing the presence of verbal agreement marks to be recovered (…)" – free translation.

The subject that is not written in the sentence are called implied subject. The implied subject of a subjectless nonfinite or verbless clause is normally identical with the subject of the superordinate clause:

(2.5) *Susan telephoned before coming over*. [...'*before Susan came over*'] (Quirk et al., 1985: 725)

As already said above, the main goal of this project is the study of subject NPs reduction. This reduction is called ellipsis in English grammars.

Quirk et al. (1985) say that ellipsis may be more strictly described as 'grammatical omission' because the omission is describable in terms of phonological units (syllables) rather than in terms of morphological units (morphemes) or grammatical units (words).

To distinguish ellipsis from other kinds of omission, they suggest defining ellipsis as a principle of *verbatim* recoverability:

> "(...) that is, the actual word(s) whose meaning is understood or implied must be recoverable. Even so, like those of so many other grammatical categories, the boundaries of ellipsis are nuclear, and it is best to recognize different degrees of 'strength' in the identification of examples of ellipsis." (Quirk et al., 1985: 884)

The criteria to be ellipsis are as follow:

a) The ellipted words are precisely recoverable;
b) The elliptical construction is grammatically 'defective';
c) The insertion of the missing words results in a grammatical sentence (with the same meaning as the original sentence);
d) The missing word(s) are textually recoverable and
e) are present in the text in exactly the same form.

(*idem*: 884-887)

All or some criteria described above can be applicable in the sentences. The authors described some subcategories according the criteria that the sentences fit. The subcategories described by the authors are:

- ✓ strict ellipsis – all five criteria apply;
- ✓ standard ellipsis – only the 'exactly copy' criterion (e) need not apply;
- ✓ situational ellipsis – this not satisfy criteria (d) and (e);
- ✓ structural ellipsis – the criteria (d) and (e) are not apply and the criterion (b) can or cannot be applied;
- ✓ semantic implication – when only the criteria (c) is applied; this is the case of the sentences that are more fittingly classified not as ellipsis at all, but as a

case of semantic implication. (*idem*: 889)

The strict ellipsis is applicable mainly to coordination. This kind of sentence can be viewed into two different ways. On the one hand the sentence can be classified as a coordination clause in which some elements of the coordination clause can be omited. On the other hand the sentence can be viewed as a single clause containing two coordinate predications[14].

The standard ellipsis can be viewed as a general textual ellipsis. There are two kinds:

i. elliptical noun phrases: there are five situations in which the noun phrase plus modifiers can be omited:

- ellipsis of postmodifier(s) alone;
- ellipsis of head + postmodifier(s);
- ellipsis of premodifier(s) + head + postmodifier(s);
- ellipsis of head alone;
- ellipsis of premodifier(s) + head

ii. elliptical clauses: the dominant type of ellipsis is final. Usually, the clause is divided into two parts: subject and operator – which remain – and predication – which is ellipted.

The situational ellipsis is dependent on the linguistic context for their interpretation. This kind of ellipsis happens more frequently in oral discourse. There two situations where ellipsis can occur: in declarative sentences and in interrogative sentences. In declarative sentences, there are the follows cases:

- Ellipsis of subject alone
- Ellipsis of subject plus operator

In interrogative sentences, there are the follows cases:

- Ellipsis of subject plus operator
- Ellipsis of operator alone

The structural ellipsis and the semantic implication might not be consider as ellipsis at all because neither the term elliptical can be a relative pronoun or it can be a case of semantic implication.

[14] In the framework of Harris, followed by this dissertation, the former perspective is adopted.

2.2 Anaphora Resolution

ZAR being a subtask of AR, first we will briefly present the literature review of AR in general and after we will present the AR for Portuguese. This AR literature review describes different approaches to resolve anaphora. Some systems are rule-base, while other focus on statistical and machine-learning approaches, including clustering algorithms.

In the literature review on AR for Portuguese, most work consist in the adaptation of an algorithm already developed for other languages, while exploring the particular features pertinent to the Portuguese Language.

As far as we know this dissertation is the first study on zero anaphora aiming at ZAR in Portuguese. The focus of this dissertation, however, is to implement a set of rules in a pre-existing system. In the future, some approaches used on the previous work can be adapted in order to develop algorithms adequate to resolve this particular type of anaphora.

2.2.1 AR using different approaches

Early works in anaphora resolution were based on linguistic knowledge and required considerable human input. Some representative works of this generation are presented below[15].

Carter (1986) shallow processing approach explored knowledge of syntax, semantics and local focusing as heavily as possible without relying on large amounts of world or domain knowledge. Carter's algorithm was restricted to nominal anaphora. His approach was implemented in a program called SPAR (Shallow Processing Anaphor Resolver). The result of this program was one of the best achieved until that time (Carter, 1986 *apud* Mitkov, 2002: 79).

Rich and Luperfoy (1988) described the pronominal anaphora resolution module of LUCY (portable English understanding system). The anaphora resolution module developed by them tried to establish coreference relations between discourse referents. There was no evaluation for this algorithm.

Carbonell and Brown (1988) proposed a general framework for intersentential anaphora resolution based on a combination of multiple knowledge sources:

[15] For this review we have tried always to consult the original papers. Whenever that was not possible, we took as main reference Mitkov (2002).

sentential syntax, case-frame semantics, dialogue structure and general knowledge. In the evaluation of this program, the success rate was 87% however this evaluation was made in a very small sample and further evaluation was considered necessary for more definitive results.

Finally, Sidner (1979) focus approach resolved full definite noun phrases and definite pronouns. Sidner assumed that a well formed discourse was about some entity mentioned in it. This entity was called the focus of the discourse (or discourse focus). According to her, there were six focus register types: discourse focus, actor focus, potential discourse focus, potential actor focus, discourse focus stack and actor focus stack. The algorithm was based on this discourse focus and was implemented in PAL (Personal Assistant Language Understanding Program) and in TDUS (Task Discourse Understanding System).

Recent works

The need to develop systems that require less linguistic knowledge and that can be applied to several languages encouraged many researches to work on knowledge-poor and robust anaphora resolution strategies.

This new strategy was facilitated through less expensaive and more reliable corpus-based NLP tools such as POS taggers and shallow parsers alongside with the increasing availability of corpora and other NLP resources. But, on the other hand, the performance of more modern approaches depends on the availability of large suitable corpora (Mitkov, 2002: 95).

Different approaches under this new paradigm are briefly described below.

Collocation patterns-based approach

Dagan and Itai (1990) described an automatic scheme for collecting statistics on co-occurrence (or collocation) patterns in a large corpus. These patterns were collected automatically from large corpora and were used to filter out unlikely candidates for antecedent.

According to the authors the use of selectional constraints presented very little success in implementing this method for broad domain. In order to avoid this low performance they proposed an alternative based on automatic acquisition of constraints from a large corpus.

As selectional constraints used in anaphora resolution require that the

antecedent satisfies the constraints imposed by the anaphor, and as this anaphor participates in a certain syntactic relation, for example being the object of some verb, then the substitution of the anaphor with the referent should also be possible since the antecedent satisfies the selectional restrictions stipulated by the verb.

Using a statistical model, the authors proposed the replacement of the candidates with the anaphor and the model would approve only those candidates which produced frequent patterns of co-occurrence.

The model had two separate phases. In the first phase, the corpus was processed and a statistical database was built. In the second phase, the statistical database is used to resolve ambiguities.

To evaluate the model, the authors used the Hansard corpus. They evaluated the reference of the anaphor 'it'. In total, they evaluated 59 sentences. The statistics were collect from part of the corpus (around 28 million words). The model proposed by the authors did not resolve 21 sentences because the threshold of 5 occurrences per alternatives could not be reached. In the remaining 38 examples the method proposed the correct antecedent 33 times (87%). Unfortunatily, results are not provided for the full set of 59 sentences.

The model proposed by Dagan and Itaí presents good results, however the problem is the need of a large corpus which most of the time is not available.

Lappin and Leass's algorithm

Lappin and Leass (1994) presented the **R**esolution of **A**naphora **P**rocedure (RAP) algorithm, which identifies the antecedents of the pronouns in intrasentential and intersentential sentences in a text. The RAP was applied with the Slot Grammar parser.

The RAP algorithm relies on measures of salience derived from syntactic structure and a simple dynamic model of attentional state to select the antecedent noun phrase (NP) of a pronoun from a list of candidates. It does not employ semantic conditions (beyond those implicit in grammatical number and gender agreement) or real-world knowledge in evaluating candidate antecedents.

During the training step, the authors used a corpus composed of five computer manuals containing approximately 82,000 words. From this corpus 560 occurrences of third person pronouns and their antecedents were extracted.

The evaluation was performed on 360 pronouns occurrences randomly selected

from a corpus of computer manuals containing 1.25 million words. RAP performed successful resolution in 86% of the cases.

The algorithm developed by Lappin and Leass presented good results. Their works is one of the most influential contributions to anaphora resolution in the 1990s: it has served as a basis for the development of other approaches and has been extensively cited in the literature (Mitkov, 2002: 105)

Kennedy and Boguraev's parse-free approach

Kennedy and Boguraev (1996) presented an algorithm for anaphora resolution which was a modified and extended version of that developed by Lappin and Leass (1994).

Once RAP algorithm operates on syntactic information alone, the authors proposed this modification because the state of the art of parsing technology still fell short of broad-coverage, robust and reliable output.

Moreover they were interested in developing a more general text-processing framework that would build its capabilities entirely on the basis of a considerably shallower linguistic analysis of the input stream, thus trading off depth of base level analysis for breadth of coverage.

Therefore the suggestion of the authors to the RAP algorithm was to work from the output of a part-of-speech tagger enriched with annotations of grammatical function. The system used a phrasal grammar for identifying NP constituents and, similarly to Lappin and Leass (1994), employed salience preference to rank candidates for antecedents.

The evaluation of this method was made with a data set containing 27 texts, taken from a random selection of genres. These texts, obtained on the basis of data from one genre only (technical manuals), contained 306 third person anaphoric pronouns of which 231 were correctly resolved, giving an accuracy of 75%, which was below Lappin and Leass's 86% accuracy. According to the authors the accuracy of this method could be improved if the tagger were more consistent regarding the gender of the words.

The modifications of the RAP algorithm enabled a larger set of text processing frameworks, with a considerably 'poorer' analysis substrate. Considering that one of the goals was to deal with a less rich level of linguistic analysis, the results showed only a small compromise in the quality of the results.

Baldwin's high-precision CogNIAC

Baldwin (1997) presented the CogNIAC (pronoun resolution program) which made use of limited knowledge and resources and its pre-processing included sentence detection, part-of-speech tagging, simple noun phrase recognition, basic semantic category information like gender, number and in one configuration, partial parse tree.

What distinguishes CogNIAC from other algorithms that use similar information is that CogNIAC does not resolve a pronoun in an ambiguous context.

Instead of using full world knowledge, CogNIAC used regularities of English use in an attempt to mimic strategies deploid by humans when resolving pronouns.

In the evaluation, the authors made two experiments. In one of them, they compared their method with Hobbs's naïve algorithm (Hobbs 1976, 1978) while the other was carried out on MUC-6 data.

In the first experiment, narrative texts about two persons of the same gender told from a third person perspective were used. Only singular third person pronouns were considered. The pre-processing consisted of part-of-speech tagging, delimitation of base noun phrases and identification of finite clauses. This pre-processing was subjected to hand correction in order to allow for the comparison with Hobbs' algorithm as far as possible.

Results, based on 298 pronouns, show 77.9% to CogNIAC against 78.8% of Hobbs' algorithm, but CogNIAC achieved higher precision (92%) even if Recall was only (64%).

For the second experiment, data from the *Wall Street Journal* were used. The performance of CogNIAC was less successful on this data with 75% precision and 73% recall. 'Software problems' accounted for 20% of the incorrect cases, another 30% were due to semantic errors like misclassification of a noun phrase into person or company, singular/plural etc. The remaining errors were due to incorrect noun phrase identification, failure to recognize pleonastic-it or other cases where there is no instance of an antecedent.

Resolution of definite descriptions

Vieira and Poesio (2000b) presented an implemented system for processing definite descriptions in arbitrary domains. The authors used definite descriptions to

indicate definite noun phrases with the definite article *the*, such as *the book*.

The system proposed by them is based on a shallow-processing approach. This system relies only on structural information, on the information provided by preexisting lexical sources such as WorldNet, on minimal amounts of general hand-coded information, or on information that could be acquired automatically from a corpus. As a result of the relatively knowledge-poor approach adopted, the system is not really equipped to handle definite descriptions which require complex reasoning; nevertheless a few heuristics have been developed for processing this class of anaphoric NPs. On the other hand, the system is domain independent and its development was based on empirical study of definite description involving human annotators.

According to Vieira and Poesio definite descriptions are separated, in the literature, into several classes but the classification schemes that they used were simpler in order to facilitate the annotation, with the purpose of getting an estimate of how well a system could do using only limited lexical and encyclopedic knowledge. Definite descriptions adopted consisted in: direct anaphora, bridging descriptions and discourse-new.

A subset of the Penn Treebank I corpus (Marcus et al., 1993) from the ACL/DCI/CD-ROM, containing newspaper articles from the *Wall Street Journal* was split in two: the first, with 1,000 definite descriptions, was used for development while the second, with 400, was kept aside by testing. The algorithm used a manually developed decision tree created on the basis of extensive evaluation.

Results on direct anaphora resolution have shown 62% of Recall, 83% of Precision and 71% of F-measure while discouse-new descriptions obtained 69% fo Recall, 72% of Precision and 70% of F-measure. Overall, the version of the system that only attempts to recognize first-mention and subsequent-mention definite descriptions obtained 53% of Recall, 76% of Precision, and 63% of F-measure. The resolution of bridging descriptions was a much more difficult task because lexical or world knowledge was often necessary for their resolutions. Around 28% of success rate in the interpretation of semantic relations between bridging descriptions using WordNet was reported.

Mitkov's anaphora resolution system

Mitkov (2002) presented the Mitkov's robust, knowledge-poor algorithm for

pronoun resolution. The algorithm used a list of preferences known as antecedent indicators.

It works from the output of a text processed by a part-of-speech tagger and an NP extractor, where it locates noun phrase candidates preceding the anaphor within a distance of two sentences[16]. It then checks candidates for gender and number agreement and, finally, it applies the indicators to the remaining candidates by assigning a positive or negative score. The noun phrase with the highest composite score is proposed as antecedent.

Mitkov's **A**naphora **R**esolution **S**ystem (MARS) is a new implementation of Mitkov's robust, knowledge-poor approach using the **F**unctional **D**ependency **G**rammar (FDG) parser as its main pre-processing tool. MARS operates in full automatic mode. In this new version, a program for automatically recognizing instances of anaphoric or pleonastic pronouns (Evans, 2000) and intrasentential syntax filter are used.

MARS operates in five phases (Mitkov, et al., 2002). In *phase 1*, the text to be processed is parsed syntactically which returns the parts of speech, morphological lemmas, syntactic functions, grammatical number and dependency relations between tokens in the text. In *phase 2*, anaphoric pronouns are identified and non-anaphoric and non-nominal instances of *it* are filtered. In *phase 3*, for each pronoun identified as anaphoric, candidates are extracted from the NPs in the heading of the selection in which the pronoun appears, and NPs in the current and preceding two sentences (if available) within the paragraph under consideration; once identified, these candidates are subjected to further morphological and syntactic tests. In *phase 4*, preferential and impeding factors are applied to the set of candidates. And, in *phase 5*, the candidate with the highest composite score is selected as the antecedent of the pronoun (Mitkov, 2002: 168).

In the evaluation, corpus of computer hardware and software technical manuals was used featuring 247,401 words and 2,263 anaphoric pronouns. Of these, 1,709 were intrasentential and 554 were intersentential. Each text was annotated for coreference relations using annotation tool CLinkA (Orasan, 2000). The overall success rate of the algorithm was 59.35%. After using a genetic algorithm (Orasan et al., 2000), the success rate rose to 61.55%.

[16] Subsequent versions have used search scopes of different lengths, 2, 3 or 4 sentences.

Mitkov's algorithm has been adapted to other languages such as Polish, Arabic, Bulgarian and Portuguese (Chaves and Rino, 2007)[17].

Anaphora resolution system for Spanish

Palomar et al. (2001) presented an algorithm for identifying noun phrase antecedents of personal pronouns, demonstrative pronouns, reflexive pronouns, and zero pronouns in Spanish. The algorithm identifies both intrasentential and intersentential antecedents and is applied to the syntactic analysis generated by the slot unification parser (SUP) (Ferrández, Palomar, and Moreno 1998b). The AR algorithm combines different forms of knowledge by distinguishing between constraints and preferences. Constraints discard some of the candidates, whereas preferences simply sort the remaining candidates[18].

In order to apply the algorithm to unrestricted texts, the authors used a partial parsing tool (Ferrández, Palomar and Moreno, 1999). This partial parse includes coordinated NPs and PPs, verbal chunks, pronouns, and what they have called free conjunctions (i.e., conjunctions that do not join coordinated NPs or PPs). Words that do not appear within these constituents are simply ignored. The NP constituents include coordinated adjectives, relative clauses, coordinated PPs, and appositives as modifiers. Sentences are then divided into clauses by parsing first the free conjunction and then the verbs.

For the identification of the type of pronoun, the authors used two approaches. In one approach, the omitted pronouns are identified with the partial-parse trees and in another approach the remaining pronouns are identified based on part-of-speech (POS) tagger out-puts.

The syntactic conditions on NP-Pronoun non-coreference are based on c-command and minimal-governing-category constraints as formulated by Reinhart (1983) and on the non-coreference conditions of Lappin and Leass (1994). In such systems, recency is important in selecting the antecedent of an anaphor, e.g. the closest NP to the anaphor has a better chance of being selected as the solution. One problem, however, is that such constraints are formulated using full parsing, whereas the main goal of this algorithm was to work with unrestricted texts (Palomar et al.,

[17] Please, see section 2.2.2.

[18] Because of some similarities with the approach used in this dissertation, this system will be presented in a little more detailed way.

2001: 550-553), a partial parsed. Therefore a set of non-coreference conditions for Spanish using partial parsing was proposed.

The algorithm was tested on both technical manuals and literary texts. A subset of the corpus Blue Book (specifically the Spanish edition of the corpus) it was selected. The Blue Book corpus consists of the handbook of the International Telecommunications Union CCITT, published in English, French, and Spanish and it contains 5,000,000 words automatically tagged by the Xerox tagger (Sánchez León, and Nieto Serrano, 1995). In the second instance, it was selected another subset from the corpus Lexesp. This corpus contains Spanish literary texts from different genres and by different authors. These texts were mainly obtained from newspapers and were automatically tagged by a different tagger than the one used to tag the Blue Book. The subset of the Lexesp corpus that was processed contained various stories, related by a narrator, and written by different authors. As was the case for the Blue Book corpus, this corpus also contained 5,000,000 words.

Both subsets selected from the Blue Book and Lexesp corpus were annotated with respect to coreference. One portion of the coreferentially tagged corpus (training corpus) was used for improving the rules for anaphora resolution (constraints and preferences), and another portion was reserved for test data.

A blind test was conducted over the entire test corpus of unrestricted Spanish texts by applying the algorithm to the partial syntactic structure generated by the slot unification parser.

Over these corpora, the algorithm attained a success rate of 76.8%. The total number of resolved pronouns was 1,677, including personal, demonstrative, reflexive, and omitted pronouns. All of them were in the third person, with a noun phrase that appeared before the anaphor as their antecedent. The “recall percentage” of the algorithm was therefore 76.8%.

According to the authors, the limitations of this algorithm are: i) mistakes in the POS tagging (causing an error rate of around 3%); ii) mistakes in the partial parsing regarding the identification of complex noun phrases (causing an error rate of around 7%) (Palomar et al., 1999); and iii) semantic information was not considered (causing an error rate of around 32%).

Machine learning approaches

Machine Learning represents learned knowledge in the form of interpretable decision trees, logical rules and stored instances. This method offers the promise of automating the acquisition of the morphology, syntax, semantic and pragmatics knowledge from annotated or unannotated language corpora by learning from a set of patterns (examples). Both decision-tree (Aone and Bennett, 1995; McCarthy and Lehnert, 1995) and instance-based methods (Cardie, 1992) have been successfully applied to resolving various types of anaphora (Mooney, 2003). Some studies that employ machine learning approaches are briefly described below.

Aone and McKee (1993) described a 'robust, extensively and manually trainable' system for multilingual anaphora resolution. They used discourse knowledge sources which where manually selected and ordered.

The continuation of that work was described in Aone and Bennett (1995, 1996) and the task was to develop truly automatically trainable systems, hoping to improve resolution performance and reduce the overhead of manually constructing and arranging such discourse data.

Their approach to build an automatically trainable anaphora resolution system consisted in tagging corpora with discourse information, and using it as training examples for a machine learning algorithm.

A corpus of Japanese newspaper articles about joint ventures has been tagged using a GUI-based tool called the Discourse Tagging Tool (DTTool) according to "The Discourse Tagging Guidelines" developed by Aone and Bennett (1994). The tool allows a user to link an anaphor with its antecedent and specify the type of the anaphor (e.g. pronouns, definite NP's, etc.). The tagged result can be written out to a SGML marked file. The tool lets the user to define types of anaphora as necessary.

The tags used for different types of anaphora were described in Aone and Bennett (1994, 1995). In this work, they also tagged the zero pronouns, a relatively common phenomenon in Japanese. For these cases, the DTTool lets the user insert a "Z" marker just before the main predicate of the zero pronouns to indicate the existence of the anaphor. The authors made distinction between QZPRO and ZPRO[19] when tagging zero pronouns. QZPRO ("quasi-zero pronoun") is chosen

[19] The authors do not provide explicit definition of the ZPRO feature.

when a sentence has multiple clauses (subordinate or coordinate), and the zero pronouns in these clauses refer back to the subject of the initial clause in the same sentence.

The anaphor types are sub-divided according to semantic criteria such as organizations, people, locations, etc. Their goal is to customize and evaluate anaphora resolution systems according to the antecedent anaphora type when necessary.

The machine learning resolver (MLR) employs the C4.5 decision-tree algorithm (Quinlan, 1993). The decision tree is trained on the basis of feature vectors for pairs of an anaphor and its possible antecedent. 66 features were used, which include lexical (e.g. category), syntactic (e.g. grammatical role), semantic (e.g. semantic class), and positional (e.g. distance between anaphor and antecedent) features.

On the training methods three parameters were used: anaphoric chains, anaphoric type identification, and confidence factors.

The training corpus used contained 1971 anaphors in 295 training texts. The evaluation corpus featured 1359 anaphors in 200 blind tests texts. Both the training and the evaluation texts were newspaper articles about joint ventures.

The evaluation was implemented on six different modes of the system. Each mode was defined on the basis of the different values of the anaphoric chain, anaphoric type identification and confidence factors. The analyses were done on the basis of only those anaphors which were identified by the program and not on the basis of all anaphors in the text.

Using the F-measure as an indicative metric for overall performance, the modes with chain parameters turned on and type identification turned off performed best with recall ranging from 67.53% to 70.20%, precision from 83.49% to 88.55% and F-measure from 76.27% to 77.27%.

McCarthy and Lehnert's RESOLVE system (1995) was created to build decision trees that can be used to classify pairs of phrases as coreferent or not coreferent. The errors generated by the sentence analyzer were eliminated by using a special tool - the Coreference Marking Interface, or CMI - to extract a set of phrases from the MUC 5 English Joint Venture (EJV) corpus (a collection of news articles, written in English, that describe business joint ventures).

In order to minimize the difficulties involved with creating and maintaining

complex sets of rules, a machine learning approach was adopted, in which a decision tree determines the order and relative weight of different pieces of evidence. RESOLVE also used the C4.5 decision tree system (Quinlan, 1993) to learn how to classify coreferent phrases.

The feature vectors used by RESOLVE were created on the basis of all pairings of reference and coreference links among them from a text manually annotated for coreferential noun phrases. The pairings that contained coreferent phrases formed positive instances, whereas those that contained noun-coreferent formed negative instances. From the 1230 feature vectors (or instances) that were created from the entity references marked in 50 texts, 322 (26%) were positive and 908 (74%) were negative.

The evaluation of the system developed by McCarthy and Lehnert (1995) focused on the coreference resolution. As all pre-processing errors were manually post-edited, the authors calculated the unpruned and pruned version of the algorithm. The results of the unpruned algorithm were: 85.4% recall, 87.6% precision and 86.5% F-measure. The results of the pruned algorithm were: 80.1% recall, 93.4% precision and 85.8% F-measure.

Soon, Ng and Lim (1999, 2001) presented a learning approach to coreference resolution of noun phrases in unrestricted text. Specifically, a coreference relation denotes an identity of reference that holds between two textual elements known as markables, which can be definite noun phrases, demonstrative noun phrases, proper names, appositives, sub-noun phrases that act as modifiers, pronouns, and so on. Thereby, according to the authors the coreference task developed by them resolves general noun phrases and is not restricted to a certain type of noun phrase such as pronouns. Also, they do not place any restriction on the possible candidate markables; that is, all markables, whether they are “organization”, “person”, or other entity types, are considered. The ability to link coreferring noun phrases both within and across sentences is critical to discourse analysis and language understanding in general.

In this system, the authors adopted a corpus-based, machine learning approach to noun phrase coreference resolution. This approach requires a relatively small corpus of training documents that have been annotated with coreferential chains of noun phrases.

A prerequisite for coreference resolution is to obtain most, if not all, of the possible markables in a raw input text. To determine the markables, a list of natural language processing (NLP) modules is used. They consist of sentence segmentation, tokenisation, morphological analysis, part-of-speech tagging, noun phrase identification, named entity recognition and semantic class determination (via WordNet).

To build a learning-based coreference engine, it is necessary to define a set of features useful in determining whether two markables corefer or not. The feature vector used in this system consists of a total of 12 features.

The machine learning algorithm used in this system is C5, which is an update version of C4.5 (Quinlan, 1993). C5 is a commonly used decision tree learning algorithm and thus it may be considered as a baseline method against which other learning algorithms can be compared.

For evaluating the system, the authors utilized the annotated corpora and scoring programs from MUC-6 and MUC-7, which assembled a set of newswire documents annotated with coreference chains. The total size of the 30 training documents is close to 12,400 words for MUC-6 and 19,000 words for MUC-7. From the MUC-6 corpus, 20,910 training examples were used, and, from the MUC-7, 48,872 training examples.

The coreference resolution system achieved a recall of 52%, precision 68%, yielding an F-measure of 58.9% for MUC-6. For MUC-7, the recall is 56.1%, the precision is 65.5%, and the balanced F-measure is 60.4%.

According to the authors their result is encouraging since it indicates that a learning approach using relatively shallow features can achieve scores comparable to those of systems built using non-learning approaches.

It should be noted that the accuracy of the coreference resolution engine depends to a large extent on the performance of the NLP modules that are executed before the coreference engine. For example the HMM named entity recognition module used by them has as score only 88.9% (considered not very high by MUC-6 standards); the part-of-speech tagger used in this system achieves 96% accuracy, while the accuracy of noun phrase identification is above 90%.

The results achieved by the coreference resolution engine cannot be directly compared with those obtained by Aone and Bennett (1995) and by McCarthy and Lehnert (1995) since these researches evaluated their systems on noun phrases that

have been correctly identified. In contrast, Soon, Ng and Lim's approach was evaluated in a fully automatic mode against the background of pre-processing errors. Also, whereas the evaluation of McCarthy and Lehnert's system was carried out on specific types of NPs (organization and business entities) and Aone and Bennett covered Japanese texts only, Soon et al.'s method processed all types of English NPs (Mitkov, 2002: 117).

Probabilistic approach

Ge, Hale and Charniak (1998) proposed a statistical method for resolution of third person anaphoric pronouns. They combined various anaphora resolution factors into a single probability which was used to track down the antecedent. The program did not rely on hand-crafted rules but instead used the Penn *Wall Street Journal* Treebank to train the probabilistic model.

In the evaluation, the data consisted of 93,931 words (3975 sentences) containing 2477 pronouns, 1371 of which were singular (*he*, *she* and *it*). The corpus was manually tagged with reference indices and referents repetition numbers. The result presented in the paper was the accuracy of the program in finding antecedents for *he*, *she*, and *it* and their various forms (e.g. *him*, *his*, *himself*, etc.) The case where *it*, i.e. the pleonastic cases, was merely a dummy subject in a cleft sentence or had conventional unspecified referents was excluded from computing the precision. They performed a ten-fold cross-validation and results are the mean success rate of all folds.

The authors investigated the relative importance of each of the above four probabilities (factors employed) in pronoun resolution. To this end, they ran the program 'incrementally', each time incorporating one more probability. Using only Hobbs's distance yielded an accuracy of 65.3%, whereas the lexical information about the gender and animacy brought the accuracy up to 75.7%, highlighting the latter factor as quite significant. The reason the accuracy using Hobbs's algorithm was lower than expected was the fact that the Penn Treebank did not feature perfect representations of Hobbs's trees. Contrary to initial expectations, knowledge about the governing constituent (co-occurrence patterns) did not make a significant contribution, only raising the accuracy to 77.9%. One possible explanation could be that selecting restrictions are not clear-cut in many cases; in addition, some of the

verbs in the corpus such as *is* and *has* were not ‘selective’ enough. Finally, counting each candidate proved to be very helpful, increasing the accuracy to 82.9%.

Based on the first experiments, the authors noted that the gender information was important making that the accuracy increases and because of that they proposed another experiment in which they considered automatic methods for estimating the probability that nouns occurring in a large corpus of English text denote inanimate, masculine or feminine things. This method is based on simply counting co-occurrences of pronouns and noun phrases, and thus can employ any method of analysis of the text stream that results in referent/pronoun pairs.

The evaluation of this new method was made with a corpus containing 21 million words of *Wall Street Journal*. The accuracy rate was 84.2%. The difference between the accuracy in the first experiment (with all factors employed) and the accuracy in the second experiment was not so high. The authors believe, however, that there are ways to improve the accuracy of the learning method and thus increase its influence on pronominal anaphora resolution.

Coreference resolution as clustering task

Cardie and Wagstaff (1999) introduce a new, unsupervised algorithm for noun phrase coreference resolution. It differs from existing methods in that it views NP coreference resolution as a clustering task. First, each noun phrase in a document is represented as a vector of attribute-value pairs. Given the feature vector for each noun phrase, the clustering algorithm coordinates the application of context-independent and context-dependent coreference constraints and preferences to partition the noun phrases into equivalence classes, one class for each real-world entity mentioned in the text. Context-independent coreference constraints and preferences are those that apply to two noun phrases in isolation. Context-dependent coreference decisions, on the other hand, consider the relationship of each noun phrase to surrounding noun phrases.

Their approach to the coreference task stemmed from the observation that each group of coreferent noun phrases defines an equivalence class. Therefore, it is natural to view the problem as one of partitioning, or clustering, the noun phrases. Intuitively, all of the noun phrases used to describe a specific concept will be “near” or related in some way, i.e. their conceptual “distance” will be small. Given a description of each noun phrase and a method for measuring the distance between

two noun phrases, a clustering algorithm can then group noun phrases together: noun phrases with distance greater than a clustering radius *r* are not placed into the same partition and so are not considered coreferent.

For the noun phrase representation, the authors follow the next steps. Given an input text, they first used the Empire noun phrase finder (Cardie and Pierce, 1998) to locate all noun phrases in the text. Next each NP in the input text was represented as a set of the features used by them. These values were automatically determined and therefore not always accurate.

The clustering approach starts at the end of the document and works backwards, comparing each noun phrase to all preceding noun phrases. If the distance between two noun phrases is less than the clustering radius *r*, then their classes are considered for possible merging. Two coreference equivalence classes can be merged unless there is any incompatible NPs in the classes to be merged.

The evaluation of the clustering approach to coreference resolution was made using the ‘dry run’ and ‘formal evaluation’ modes (MUC-6). For the ‘dry run’ data set, the clustering algorithm obtained 48.8% recall and 57.4% precision, which came to an F-measure of 52.8%. The formal evaluation scores were 52.7% recall and 54.6% precision, coming to an F-measure of 53.6%. Both runs used $r = 4$ which was obtained by testing different values on the dry run corpus. Different values of *r* ranging from 1.0 and 10.0 were tested and, as expected, the increase of *r* raised recall, but lowered precision.

The clustering approach was also compared with three baseline algorithm. The first baseline marked every pair of noun phrases as coreferent, i.e. all NPs in the document form one class, scoring 44.8% F-measure for the dry run data test and 41.5% for the formal run dataset. This baseline is useful because it establishes an upper bound for recall on clustering algorithm (67% for the dryrun and 69% for the formal evaluation). The second baseline considered each two NPs that have a word in common as coreferential; it produced scores of 44.1% and 41.3% respectively. Finally, the third baseline marked as coreferential only NPs whose heads matched; this baseline obtained F-measures of 46.5% and 45.7% respectively.

The limitations of the Cardie and Wagstaff’s approach arise from the greedy nature of the algorithm and in the low accuracy of the pre-processing: NPs are identified at base level only; most of the heuristics for computing the 11 features are very crude.

2.2.2 AR for Portuguese

In Portuguese, there are not so many works in anaphora resolution such as in English. In this section a selection of the most recent works is presented.

Coelho (2005) presented an adaptation of the Lappin and Leass's (1994) algorithm for Portuguese. The proposed algorithm has all the main components of the original algorithm, with the following differences: i) the syntactic filter and the anaphor binding algorithm were replaced for the coreferential restrictions proposed by Reinhart (1983); ii) the grammar parser used was PALAVRAS (Bick, 2000); iii) the Xtractor (Gasperin et al*.,* 2003) tool was used to convert the grammar parser output in XML; iv) the procedure for identifying the pleonastic pronouns *it* was not implemented because Portuguese does not have such cases; and v) the cataphora phenomenon was not considered.

In the evaluation of the algorithm, three corpora were used: legal corpus, literary corpus, and journalistic corpus. The legal texts' corpus was composed with legal opinion of the Attorney-General of the Republic of Portugal. The literary corpus was composed by the book *O Alienista* by Machado de Assis. And the journalist corpus was composed with 14 journalist texts.

All corpora were automatically annotated by PALAVRAS with morphological and syntactical information; person pronouns were manually annotated using the MMAX (*Multi-Modal Annotation in XML*) tool.

The evaluation was made in three experiments. In the first experiment, the legal texts' corpus was used. The solution generated by the algorithm was considered correct when it was the same as the solution annotated manually or when the NP generated contained the NP annotated manually. The results of the algorithm were: 35.15% anaphora correctly resolved and 63.8% anaphora poorly resolved[20].

The second experiment was made using the literary corpus. The criterion to check if the solutions were or not correct was the same as in the first experiment. The results were: 31.32% anaphora correctly resolved and 68.68% anaphora with wrong solution.

In the third experiment, the literary corpus and the journalist corpus were used. The literary corpus was processed over again because an error occurred in the

[20] The algorithm chose the antecedent erroneously.

morphological and syntactical information about the gender of the words; to correct this problem, a manually annotation was made in the literary texts. Otherwise, the same procedure was adopted in the other experiments. The results were: 32.61% (literary corpus) and 43.56% (journalistic corpus) anaphora correctly resolved; 67.39% (literary corpus) and 56.44% (journalistic corpus) anaphora with wrong solution.

The results obtained for the algorithm adapted for Portuguese presented a smaller score when compared with the results obtained for the original algorithm. This happened in part because the original algorithm was evaluated using computer science textebooks and manual and the adaptation of the algorithm used texts of different genre. Besides, 46.84% of the pronominal anaphora was composed of the pronoun *lhe(s)* and *se* whose NPs antecedents can be masculine or feminine, making the resolution of these pronouns more complex since the morphological filter does not eliminate any of the NP candidates.

Another problem was that the parser PALAVRAS assigned morphological and syntactical incorrect information and incorrect identification of the reflexive and reciprocal pronouns.

The XML file generated by the Xtractor presented some problems also. Some information in the PALAVRAS's output was not processed damaging the final results.

Finally, the salience weights were optimized for the English and those rates should have been reviewed for Portuguese.

Chaves and Rino (2007) presented an adaptation of the Mitkov's algorithm[21] for Portuguese. The RAPM (***R**esolução **A**nafórica do **P**ortuguês baseada no algorítmo de **M**itkov* 'Anaphora Resolution for Portuguese based on Mitkov's algorithm') differs from the original algorithm in that it aims the Brazilian Portuguese and its input texts were automatically annotated unlike the Mitkov's approach in which the morphosyntactic annotations were manually corrected before going into anaphora resolution (Chaves and Rino, 2007:53).

Moreover, to resolve morphological dependencies, RAPM looks up an XML onomastic file with correct information of gender and number of the proper nouns, and the antecedent search scope is of three sentences, instead of two. The XML file

[21] This adaptation was based on the version presented in Mitkov (1998).

with the proper nouns extracted from a text corpus was used to minimize preprocessing problems. In the absence of such information, they would be assigned both genders and numbers. The last distinction from the original algorithm is that at this time RAPM did not incorporate modules for preprocessing.

Unlike the original algorithm which used eleven antecedent indicators, the adaptation of the algorithm used only five and three other new indicators were added. The antecedent indicators were: i) *First NP (FNP)*; ii) *Lexical Reiteration (LR)*; iii) *Indefinite NP (INP)*; iv) *Prepositional NP (PNP)*; v) *Referential Distance (RD)*; vi) *Syntactic Parallelism (SP)*; viii) *Nearest NP (NNP)*; and ix) *Proper Noun (PN)*.

RAPM was evaluated using success rate as the evaluation measure No correction procedure was applied to the input data, aiming at a more realistic black-box approach in the future.

The corpora used for the evaluation were the same used in Coelho (2005): legal texts (with 110, 610 words), literary texts (with 16,530 words), and journalistic texts (with 13,217 words).

The data files used were automatically annotated by Coelho. Such input was produced in the following way: raw texts were parsed by PALAVRAS and converted to XML by the Xtractor tool.

The evaluation was done using different combinations of the antecedent indicators[22] when running on the journalistic corpus. The best performance was using the system discriminated as RAPM_8 – 67.01% of success rate.

Then, the strategy with the best success rate (RAPM_8) was used in another experiment: the results were compared with two distinct baselines, namely, 'Baseline-NP'[23] and 'Baseline_Subj'[24] – the same baselines used by Mitkov (2002). Chaves and Rino system scored 67.01%, Baseline-NP scored 55.49% and Baseline_Subj scored 42.27% of success rate.

Comparing the success rates of RAPM systems with Coelho's system, the RAPM system was consistently superior regarding the three corpora. And the comparison with the baseline scores showed that the system presented an improvement in pronominal anaphora resolution for Portuguese.

[22] For more details on the different combinations, please see Chaves and Rino (2007).

[23] Baseline-NP checks agreement in number and gender and, when more than one candidates remains, picks out as antecedent the most recent noun phrase matching the gender and number of the anaphor.

[24] Baseline_Subj adds to the Baseline-NP a third constraint: the antecedent NP must occupy the subject position in the sentence it occurs.

Santos (2008) presented an adaptation of the Hobb's algorithm for Portuguese. The author chose to use only syntactical information in order to discover how important the syntactic information is to the resolution of referential pronouns in Portuguese.

The original algorithm did not resolve reflexive pronouns, but in the Portuguese adaptation, the authors included it.

The corpora used in the evaluation were composed by the corpora utilized by Coelho (2005) plus the corpus Summ-it (Collovini et al., 2007). All corpora were processed by the parser PALAVRAS and the Xtractor tool and were then manually post-edited to ensure that the input of the algorithm was correct. The legal texts' corpus, literary corpus and journalistic corpus were already described above. The Summ-it corpus is composed by 50 journalistic texts from the science section of the newspaper *Folha do Estado de São Paulo*. In this corpus, the coreference was manually annotated.

In the journalistic, literary, and Summ-it corpora, the solution proposed by the algorithm was considered correct if the referent was the same that it was annotated in the corpus or if the generated solution was coreferent of the solution annotated. However, in the legal texts corpus, the solution was considered correct if the solution generated was the same of the annotated solution.

The system scored 52.45% of success rate for the reflexive pronouns. For the non reflexive pronouns it scored 44.48% and, in general, the success rate was 45.84%.

Comparing the results presented above with the results of the Coelho (2005) adaptation, one can conclude that both algorithms had an equivalent performance regarding the non-reflexive pronouns; however Santos (2008) algorithm has succeeded and contributed to improvement of the performance of the algorithm regarding the reflexive pronouns.

The work of Cuevas et al. (2008) focuses on pronoun resolution as required by Portuguese-Spanish-English MT project under development. Their present choice of target – Portuguese third person plural pronouns (*Eles*[25]/*Elas*) – is based on the assumption that these pronouns are less prone to ambiguity, and arguably easier to

[25] Notice, however, that in Portuguese *eles* 'they_ms_pl' can also present an indefinite reading.

resolve than the English equivalent (*They*), which may suggest an interesting multilingual approach to anaphora resolution.

As a first step to boost translation performance in these languages, some basic resources for Portuguese was built, namely, a coreference annotation tool, an annotated corpus and training data derived from tagged text. Secondly, the usefulness of this preliminary data was evaluated in two standard machine learning approaches to pronoun resolution (statistical/unsupervised and symbolic/supervised).

The corpus used in this procedure is composed by 646 articles (440.690 words in total) from the Environment, Science, Humanities, Politics and Technology supplements of the on-line edition of the *Revista Pesquisa FAPESP*, a Brazilian journal on scientific news. The resulting corpus was tagged using the PALAVRAS tool (Bick 2000).

As it was said, for this study on anaphora resolution, only third person plural pronouns *eles* (masculine) and *elas* (feminine), which are both translated as (no gender-specific) *they* in English. 813 instances of such pronouns (584 masculine and 229 feminine) were found in our corpus.

In order to take advantage of the (Portuguese) information made available by PALAVRAS, a simple coreference annotation tool from scratch was developed. Besides providing the basis for the training data, the use of the existing tags allowed to constrain automatically the choices to be made by the human annotator regarding both referring expressions (which are user-defined) and potential antecedents (taken to be the existing NPs, etc.).

Two independent annotators used the tool to link each of the selected instances of reference to their antecedents in the text, except for the cases of reference to compound antecedents (e.g., *John and Mary*) which were not presently addressed.

Following the annotation task, the annotators compared their data and excluded all instances of reference on which they could not immediately reach agreement. This was mainly the case of errors introduced by the tagger itself (i.e., unidentified NPs) and ill-formed or ungrammatical sentences. As a result, the set of 483 revised instances of reference to single terms was selected. This data set was the basis of the training data.

The authors based on the work of Soon et al. (2001) to perform this task. The present pronoun resolution task was considered as a classification problem in which a pronoun *p* and a potential antecedent *a* may corefer or not. To this end, it was

considered positive instances of coreference the pairs (*p*, *a*) explicitly defined as coreferential by the annotators, and it was considered negative instances all pairs (*p*, *a*) in which *a* is an intermediate NP between *p* and its actual antecedent.

The first experiment was based on an unsupervised statistical approach, the EICAMM (Enhanced ICA Mixture Model) (Oliveira and Romero, 2004), which is an extension of the ICA Mixture Model (ICAMM) (Lee et al., 2000). Using the entire set of features, the algorithm correctly classified 1797 (76.11%) instances. Regarding the coreferential class, the algorithm scored 43.1% of precision, 93.1% of recall and 59.0% of F-measure. In the non coreferential class it was scored 97.9% of precision, 72.2% of recall and 83.1% of F-measure.

The second experiment involved the induction of decision trees. Using ten-fold cross-validation and all the features, in the coreferential class the algorithm obtained 67.9% of precision, 52.0% of recall and 58.9% of F-measure and in the non coreferential class, the results were 89.7% of precision, 94.4% of recall and 92.0% of F-measure.

These results suggest that – at least for this data set – there was no useful relation between the syntactic position of the pronoun and its antecedent. However, the low precision levels for coreferential cases indicate that additional features (possibly making use of semantic knowledge) are indeed required.

In order to improve the results achieved in the experiments above (Cuevas et al. 2008) especially to improve the precision measure for the coreferential cases, Cuevas and Paraboni (2008) proposed to extend the set of the features including several features intended to capture syntactic constraints that are central to pronoun resolution, besides additional semantic information required to disambiguate cases of coreference in which there is no number agreement between pronoun and antecedent (e.g., “The company” and “They”.). At the same time, the same general principle of limiting the feature set to the kind of knowledge available from the PALAVRAS tag set was kept.

Apart from the extended set of features, a more comprehensive evaluation work in a second linguistic domain, and an initial attempt to cover singular instances of pronouns, which were not originally included in the training data was also presented. In addition to that, as the current approach reached satisfactory success rates, for

the first time the original test data left aside in the previous work was used. To this end, 13 learning features (plus the coref class to be learned) were considered.

For the evaluation, the corpus used in this experiment was the same used in Cuevas et al. (2008). The main test (test 1) consisted of a standard C4.5 decision-tree induction approach (Quinlan, 1993). The results were: in the coreferential class, 85.7% of precision, 87.5% of recall and 86.6% of F-measure and in the non coreferential class, 96.3% of precision, 95.7% of recall and 96.0% of F-measure.

As a second test (test 2) a different corpus was used, namely articles taken from the 1994 politics supplements of the *Folha de São Paulo* newspaper. However, as the time was not enough to build the required (and necessarily large) training data in the new domain, it was decided to verify how much loss in accuracy the existing model (trained on science magazines) would experience if applied to the resolution of pronouns found in newspapers. The results were: in the coreferential class, 68.3% of precision, 69.1% of recall and 68.7% of F-measure and in the non coreferential class, 93.7% of precision, 93.4% of recall and 93.6% of F-measure.

Finally in the third tests the entire (and hence mixed) data set was used (2603 instances in the science magazines domain and 477 instances in the newspapers domain), and along with a ten-fold cross validation. The results were: in the coreferential class, 72.4% of precision, 70.3% of recall and 71.3% of F-measure and in the non coreferential class, 93.4% of precision, 94.0% of recall and 93.7% of F-measure.

Despite the still insufficient amount of training instances in the newspapers domain (recall that the amount of instances from the science domain is over six times larger) the results show considerable improvement, with an average 89.64% correctly classified instances (71.3% F-measure in coreferential cases). The results of this investigation show major improvement in resolution accuracy over the previous work (Cuevas et al., 2008).

2.3 Zero Anaphora Resolution

The literature review of ZAR will be presented below focusing on studies for Japanese, Chinese, Spanish and Portuguese.

2.3.1 ZAR for Japanese

There many studies in zero anaphora resolution for Japanese. In this dissertation only some recent works were selected.

Seki et al. (2001) proposed a method to resolve Japanese zero pronouns which uses a probabilistic model decomposed into syntactic and semantic properties. The syntactic model was trained based on corpora annotated with anaphoric relations, and the semantic model was trained based on a large-scale unannotated corpus, so as to counter the data sparseness problem. In this work, solely zero pronouns whose antecedents exist in preceding sentences have been focused since they are major reference in Japanese discourse.

The process of the Japanese Zero pronoun resolution proposed by Seki et al. (2001) is performed the following steps: 1) given as input Japanese texts, the system performs the morphological and syntactic analyses; 2) the zero pronoun identification is made through the case frame dictionary; 3) in the zero pronoun resolution phase, the antecedent candidates for each zero pronoun are extracted from the text using the syntactic model (which was trained based on annotated corpora) and the semantic model (which was trained based on unannotated corpus). Based on previous experiences in zero pronoun resolution, the authors used six features.

According to the authors, the system developed by them was made to be contextualized as a module in NLP applications, such as machine translation systems. In those applications, it is desirable that the resolution module selectively outputs antecedents that are resolved with a higher certainty degree, so as to improve the accuracy of the system (consequently, the system coverage potentially decreases). Thinking on this problem, the notion of certainty was introduced in the probabilistic model. It is assumed that system outputs (i.e., antecedents with the greatest probability score) are more likely to be correct in the following two cases: i) the probability score for the first antecedent is sufficiently great and ii) the probability score for the first antecedent is significantly greater than that for the second antecedent candidate.

In the evaluation the Kyotodaigaku Text Corpus version 2.0 was used, in which 20,000 articles included in Mainichi Shimbun newspaper published in 1995 were analyzed by a morph/syntax analyzers and manually revised. From this corpus, a

random sampled 30 editorials and 30 general articles (e.g., politics and sports) were selected. Editorials were distinguished from other articles because, i) they are mainly subjective opinions while general articles are relatively objective and, ii) this difference potentially affects zero pronoun resolution. The sample articles were annotated manually with anaphoric relations. Accuracy was adopted as the evaluation metrics.

In the evaluation two models were compared: 1) the probabilistic model using all features (*both2*) and 2) the control (baseline) model, which adopted the following rules: a) semantic consistency between a zero pronoun and its antecedent candidate, b) proximity between a zero pronoun and its antecedent candidate, c) a post-positional particle that follows an antecedent candidate.

The probabilistic model (*both2*) was tested three times and the accuracy was: in ranking 1, 39.8% for the editorial corpus and 54.0% for the general corpus; in ranking 2, 55.2% for the editorial corpus and 66.2% for the general corpus; and in ranking 3, 62.4% for the editorial corpus and 75.5% for the general corpus.

For the baseline model (*rule*) the procedure was the same (it was tested three times) and the accuracy was: in ranking 1, 36.1% for the editorial corpus and 38.9% for the general corpus; in ranking 2, 52.0% for the editorial corpus and 52.1% for the general corpus; and in ranking 3, 59.2% for the editorial corpus and 62.5% for the general corpus.

The accuracy related to editorials was lower than one for general articles. This result implies that the domain of an input text affects the accuracy of Japanese zero pronoun resolution. Furthermore the *both2* model outperformed the *rule* model. Thus, the conclusion is that the model integrating syntactic and semantic information was effective for zero pronoun resolution.

Isozaki and Hirao (2003) proposed a method that combines ranking rules and machine learning. Heuristic ranking rules give a general preference, while a machine learning method excludes inappropriate antecedent candidates.

The corpus used by them was the same used in Seki et al. (2001). It was made some adjustments in the corpus like ambiguous antecedents which was replaced by the explicit names and it was removed zero anaphors in quoted sentences.

The authors decided to use the output of ChaSen and CaboCha instead of the morphological information and the dependency information provided by the Kyoto

Corpus since the classification of the joshi (particles) in the Corpus was not satisfactory for their purpose.

In the evaluation the authors used different combinations and the best result was 66.3% zero anaphors correctly resolved for the general corpus and 50.2% for the editorial corpus.

According to the authors it is not possible compare their results with the Seki's results because the data used in this experiments was slightly different from Seki's.

Iida et al. (2007) proposed a method to resolve zero-anaphora by decomposing it into intrasentential and intersentential zero-anaphora resolution tasks. According to them, for the intrasentential task, syntactic patterns of zero pronouns and their antecedents are useful information. The authors considered only zero-pronouns that function as an obligatory argument of a predicate for this work.

The method adopted by the authors consisted of use the Japanese morphological analyzer ChaSen and the dependency structure analyzer CaboCha, which also carries out named-entity chunking, to obtain the dependency parse tree, in which words are structured according to the dependency relation defined in the Kyoto Corpus. Then it was extracted the path between a zero-pronoun and its antecedent. Finally, to encode the order of siblings and reduce data sparseness, the authors transformed the extracted path.

The learning algorithm selected was the BACT system. This system learns a list of weighted decision stamp with a boosting algorithm. Each decision stamp classifier is represented as a labeled ordered tree appearing in the training instances. In the proposed anaphoric determination problem, given a set of positive (anaphoric) training trees and a set of negative (no anaphoric) training trees, BACT induces a set of sub trees (decision stumps) that are useful for the binary classification. The BACT algorithm has the important characteristic that the results of learning trees are more human readable, because the result of each iteration is given as a pair of decision stumps and weight.

For the evaluation, the authors used Japanese newspaper articles. The data set contained 1,384 intrasentential anaphoric zero-pronouns, 1,128 intersentential anaphoric zero-pronouns, and 784 non-anaphoric zero-pronouns (3,306 zero-pronouns in total), with each anaphoric zero-pronoun annotated to be linked to its antecedent. For each experiment, it was used 137 articles for training, 60 articles for

optimizing (threshold parameter of intrasentential zero-anaphora resolution), and 150 articles for testing.

The authors tested different combinations of the features and the best result was using syntactic patterns features: 59.6% of recall, 59.5% of precision and 59.5% of F-measure. Taking the Japanese as a target language, it was empirically demonstrated that incorporating rich syntactic pattern features in a state of the art learning-based anaphora resolution model dramatically improved the accuracy of intrasentential zero-anaphora, which consequently improved the overall performance of zero-anaphora resolution.

Sasano et al. (2008) presented a probabilistic model for Japanese zero anaphora resolution. First, this model recognizes discourse entities and links all mentions to them. Zero pronouns are then detected by case structure analysis based on automatically constructed case frames. Their appropriate antecedents are selected from the entities with high salience scores, based on the case frames and several preferences on the relation between a zero pronoun and an antecedent. Case structure and zero anaphora relation are simultaneously determined based on probabilistic evaluation metrics.

To training the probabilistic model and to evaluate the proposed model, the authors created an anaphoric relation-tagged corpus consisting of 186 web documents (979 sentences). It was selected 20 documents for test and used the other 166 documents for calculating several probabilities. In the 20 documents, 122 zero anaphora relations were tagged between one of the mentions of the antecedent and the target predicate that had the zero pronouns.

Each parameter for proposed model was estimated using maximum likelihood from the data. The case frames were automatically constructed from web corpus comprising 1.6 billion sentences. And the case structure analysis was conducted on 80 million sentences in the web corpus.

The authors annotated manually the morphemes, named entities, syntactic structures and coreferential relations. Since correct coreferential relations were given, the number of created entities was the same between the gold standard and the system output.

For the proposed task, the results were: 42.6% of recall, 27.1% of precision and 33.1% of F-measure. According to the author, it is needed to improve the system and

as a future work, they plan to conduct large-scale experiments and integrate this model to a fully lexicalized probabilistic model for Japanese syntactic and case structure analysis.

As we can see, the recent works for Japanese use hybrid system in which the linguistic analyses and the machine learning or probabilistic model are implemented together. The framework adopted for these works is different than the framework proposed in this dissertation.

2.3.2 ZAR for Chinese

Yeh and Chen (2003) proposed resolve zero anaphora that occur in the subject position or in the object position. Their approach relies on limited knowledge and only need partial syntactic parsing of text. The resolution process works from the output of a POS tagger enriched with annotations of grammatical function of lexical items in the input text stream. The partial parsing technique is used to detect zero anaphors and identifies the noun phrases preceding the anaphors as antecedents. The authors also employ centering theory and constraint rules to identify the antecedents of zero anaphors appeared in the preceding utterances.

The ZA resolution method is divided into three parts. First, it was used a POS tagger to produce the tagged result of an input document. Second, the ZA is detected by employing detection rules based on the result of partial parsing. Third, the antecedent of the ZA is identified using rules based on the centering theory.

For the evaluation, the authors made two experiments. In the first experiment, during the ZA detection phase, they employed the ZA detection rules as the baseline obtaining 65.2% of precision. Then they added the ZA detection constraint to see the result and the precision obtained was 80.5%.

In the second experiment, during the antecedent identification phase, it was used the rule without involve the centering theory. The results were: 65.8% of recall and 55.3% of precision. Then it was used the rule with the centering theory in order to compare the improvement of their method. The results were: 70% of recall and 60.3% of precision.

The test corpus was composed by a collection of 150 news articles contained 998 paragraphs, 4631 utterances, and 40884 Chinese words.

Some errors detected by the authors were: i) when the antecedent of the zero anaphora was in the preceding utterance, ii) when a ZA referred to an antecedent mentioned in the succeeding utterances (cataphora), and iii) when the ZA resolution depended on the background knowledge of readers.

Peng and Araki (2007) proposed a learning classifier based on maximum entropy (ME) for resolving zero-anaphora in Chinese text. This work focused only on the identification of the antecedent of the ZA because, according to the authors, the task of the ZA anaphora detection can be performed by some other modules such as a shallow parser.

They constructed a maximum entropy (ME) classifier to check whether a candidate is the correct candidate or not. First they employed a set of 13 regular features to capture the context information in discourses. But they noticed that it was needed to improve the classifier especially because the semantic knowledge was insufficient. Thus, they developed the Web-based features to obtain additional semantic information from the Web.

To evaluate the importance of the Web-based features, the authors tested the system with the 13 regular features – which scored accuracy of 71.9% – and, after, they tested the system with the Web-based feature – which scored the accuracy of 81.8%.

According to the authors, as these experiments showed promising results, the Web-based feature can be effectively introduced into the machine learning framework and thus increase the performance of the ZA resolution.

Wu and Liang (2009) proposed a new approach for ZA resolution applying case-based reasoning (CBR) and pattern conceptualization. According to the authors the CBR is able to exploit the previous experience that might be useful for the problem. For this experiment, the authors utilized the antecedent features of the retrieved cases to predict the antecedent of a new case. As all cases were represented with the patterns containing semantic tags for their nouns and grammatical tags for the verbs, such pattern conceptualization will be able to efficiently reduce data sparseness in the case base. Moreover, the presented resolution was incorporated with a filtering mechanism to identify those non-

anaphoric cases such as cataphora and non-antecedent instances in order to enhance the overall resolution performance.

For the evaluation of the proposed approach, the authors utilized 382 narrative articles selected from ASBC corpus. They used the fivefold cross-validation over the selected data set. The experimental results showed that the proposed approach achieved good results by yielding 79% of F-measure, on 1,051 instances of ZA.

As we can see, the recent works on ZA resolution for Chinese follows the same premise of the researches developed for the Japanese. As stated above, the framework adopted for these works is different than the framework proposed in this dissertation.

2.3.3 ZAR for Spanish

Ferrández and Peral (2000) proposed a computational approach for resolving zero pronouns in Spanish. The authors worked only with zero pronouns that appeared specifically on the subject position. The resolution of these pronouns was implemented in the computational system called Slot Unification Parser for Anaphora resolution (SUPAR) (Ferrández et al., 1999).

The ZA resolution proposed by the authors was based on the distinction between preference and restriction heuristics which employed information originating from morphosyntactic or shallow semantic analysis. The authors used a partial parsing to detect the zero pronouns and to give the necessary information for the preference and restriction heuristics used in this system. The number of previous sentences considered to select the antecedent of the zero pronouns was four sentences.

To training the system, the authors used a handmade corpus which contained 106 zero pronouns. This corpus was used mainly to define the order of the preference heuristic. To evaluate the system, it was made a blind test on unrestricted texts. Specifically, SUPAR has been run on two different Spanish corpora: i) a part of the Spanish version of *The Blue Book* corpus, which contains the handbook of the International Telecommunications Union CCITT, published in English, French and Spanish, and automatically tagged by the Xerox tagger, and ii) a part of the *Lexesp* corpus, which contains Spanish texts from different genres and authors. In general, the system achieved 75% of success rate.

The work on ZA resolution described above for Spanish is based on preference and constraints heuristics like some works on AR including the Mitkov´s approach. Once again, the framework adopted is different than the framework proposed in this dissertation.

2.3.4 ZAR for Portuguese

Carvalho and Madura (2002) developed and implemented a syntactically-based algorithm that recovers the omitted constituents and reconstructs the elliptical clause, when applicable. This algorithm deals only with sentences involving coordination and ellipsis simultaneously and takes *Island Constraints* into account in order to reconstruct the omitted material.

The basic strategy which the algorithm was encoded is to reconstruct the omitted clause by i) decomposing the sentence into syntactic structures; ii) identifying the type of ellipsis present in the sentence; iii) checking if this type of ellipsis is subject to syntactic constraints; iv) identifying the antecedent; and v) reconstructing the omitted constituent. The evaluation of this algorithm is not available.

Again the framework adopted on this work is not the same framework adopted on this dissertation.

3 Scope and Methods

This dissertation focuses on the construction of linguistically motivated rules for zero anaphora resolution, to be integrated in the XIP parser. Because of the complexity of the subject, some delimitation of this general objective has to be made. Section 3.1 further specifies the scope of the dissertation, while section 3.2 presents in detail the sentence types here addressed, and in which the zeroing of the subject NP occurs. Section 3.3 briefly presents the methods here used and in 3.4 a comprehensive description of the corpus is provided. Finally, in the last section (3.5), linguistically motivated rules, implemented on the XIP parser are presented and justified.

3.1 Scope

Based on the linguistic knowledge of Portuguese and on the preliminary results of the corpus described below, we define as follows the scope of this dissertation:

a) only subject NP deletion will be considered;
b) NP deletion will only be solved within sentence boundaries and with an explicit antecedent;
c) rules are to be formalized based solely on the results of the shallow parser (or chunks), that is, with minimal syntactic (and no semantic) knowledge;
d) other restrictions on scope will also have to be made, and we will present them in the appropriate place.

3.2 Sentence types

Zeroed NP subjects are non-explicit (hidden) subjects in complex sentences: coordinative and subordinate sentences.

3.2.1 Coordinate sentences

A clause is classified as coordinate when it does not have a syntactic (argument-like or adverbial) function in relation to another clause. Beside the second clause and the coordinating conjunctions cannot be fronted (Matos, 2003):

(3.1) *João e Maria*$_{i}$ *viajaram para o Sul mas Ø*$_{i}$ *não foram de férias*

"John and Mary travelled to the South, but they were not in vacation"

**Mas não foram de férias, João e Maria viajaram para o Sul*

*'But they were not in vacation, John and Mary travelled to the South'

The main element in coordinate sentences is the coordinating conjunctions whose function is to make explicit the relation between the coordinated terms (*idem*: 558).

Coordinate cluases are also divided into two types: syndetic and asyndetic. The difference between them is the use of the conjunction: while in the coordinative syndetic sentence the conjunction is expressed (3.2), in the coordinative asyndetic sentence it is not expressed (3.3).

(3.2) *Às vezes ele*$_{i}$ *atrasa o pagamento das contas mas, depois, Ø*$_{i}$ *paga*

'Sometimes he delays the payment of the bills but, after, [he] pays'

(3.3) *O João*$_{i}$ *acordou, Ø*$_{i}$ *escovou os dentes*

'John waked up, [he] brushed his teeth'

The coordinative syndetic clauses have three subtypes: additive, adversative and alternative. The conjunctions used for each type are described in Appendix 1. Discontinuous morphemes like conjunctions *não só… mas também* 'not only… but also' otherwise equivalent to the additive *e* 'and', will not be considered in this study.

3.2.2 Subordinate sentence

Subordinate clauses has a syntactic (argument-like) function in relation to the main clause (nominal subordinate clause) (3.4); or modifies a head noun being part of its NP (relative clauses) (3.5); or it expresses circumstantial events that modify the main clause (adverbial subordinate clause) (3.6).

(3.4) O João disse que não estava se sentindo bem

'John said that [he] was not feeling good'

(3.5) A Maria, que vestia uma roupa vermelha, foi ao funeral do marido

'Maria, who was wearing a red suit, went to the funeral of her husband'

(3.6) O tempo mudou quando anoiteceu

'The weather has changed when it got dark'

In this dissertation, only nominal subordinate clause and adverbial subordinate clause will be dealt; the adjective subordinate clause will not be considered because the relative pronoun may or not exercises the syntactic function of subject in the

sentence and, at this time, it is not possible for the grammar to discriminate in which cases the relative pronoun is the subject.

3.2.3 Nominal subordinate clause

Nominal subordinate clause can be finite (the verb is in the indicative or subjunctive mode) or non-finite (the verb is in the infinitive).

Finite nominal subordinate clause are introduced by the conjunctions *que* 'that' or *se* 'if'. The conjunction *se* 'if' is usually used when the verb of the main clause is an **inquire verb** like *investigar* 'to investigate', *perguntar* 'to request', or a **doubting verb** considerably negative like *desconhecer* 'to not know', *ignorar* 'to ignore', *não saber* 'to not know'. Besides when the verb of the main clauses clause is **declarative verb** like *decidir* 'to decide', *descobrir* 'to discover', *dizer* 'to say', *informar* 'to inform', *mostrar* 'to show' the conjunction used can be *se* 'if' or *que* 'that'. These conjunctions are called integrant conjunction.

The nominal subordinate clause in which the NP subject can be zeroed is divided into three types[26]. This division is made based on the syntactic function that the subordinate clause exercises regarding the main clause. The three types are:

a) A clause acting as the subject of the main clause:

(3.7) *Não é preciso que as prestações*$_i$ *sejam do mesmo valor, basta que Ø*$_i$ *sejam da mesma natureza*

"It is not necessary that the installments are the same value; it is enough that [they] are similar"

b) A clause acting as direct (accusative) object of the main clause:

(3.8) *Ele*$_i$, *naquele momento, tinha dúvidas: não sabia se Ø*$_i$ *ia à praça e enfrentava o povo, ou se fugia para longe*

"He, in that moment, had doubts: [he] did not know if [he] went to the square and faced the people or if [he] fled away"

(3.9) *Os primos*$_i$ *acham que Ø*$_i$ *estarão usando a coleção daqui a quarto ou cinco anos*

"Cousins think that [they] will be using the collection from now or five years"

c) A clause acting as indirect object of the main clause:

[26] Grammars also consider appositive clauses as a case of subordintation. Following Harris (1991), we prefer to integrate apposition as a form of coordination. Still, apposition was not considered in this study because of the small number of cases found in the corpus.

(3.10) *Fleury*$_i$ *insiste em que* Ø$_i$ *apenas deu despachos interlocutórios*

"Fleury insists that [he] only gave interlocutory orders"

(3.11) *No Palmeiras, todos*$_i$ *estão conscientes de que hoje* Ø$_i$ *têm um grande desafio pela frente*

"At Palmeiras, everyone is aware that today [everyone] has a great challenge ahead"

In the non-finite nominal subordinate clause, the integrant conjunction is not used and the verb of the subordinate clause is in the infinitive. The three types described above are the same:

a) A clause acting as subject of the main clause:

(3.12) *Maria*$_i$ *disse que é urgente* Ø$_i$ *partir imediatamente*

"Mary said that it is urgent to depart immediately"

b) A clause acting as direct object (accusative) of the main clause:

(3.13) *O Zezé*$_i$ *disse* Ø$_i$ *ter matado o sindicalista Oswaldo Cruz Júnior em legítima defesa*

Zeze said to have killed the union leader Oswaldo Cruz Jr. in self-defense

c) A clause acting as indirect object (dative cases) of the main clause:

(3.14) *O João*$_i$ *recorda-se de* Ø$_i$ *ter sido campeão*

"John remembers being champion"

(3.15) *Paula*$_i$ *estava ansiosa para* Ø$_i$ *voltar*

"Paula was anxious to come back"

3.2.4 Adverbial subordinate clause

Adverbial subordinate clauses are characterized by exercising the syntactic function of adverb in relation to the main clause.

In the finite construction of the adverbial subordinate clause, the conjunction is used and the verb is in the indicative or subjunctive mode. In the non-finite construction, the conjunction is sometimes optional and the verb is in the infinitive, the gerund or in the past participle. The (non exhaustive) list of the conjunctions considered in this dissertation is provided in Appendix 1.

The adverbial subordinate clause in which the NP subject can be zeroed is divided into six types[27]. This division is made based on the semantic information of the adverbial subordinate clause. The types are:

[27] Some grammars consider nine types, but, in this dissertation, the comparative, the conformative and the proportional adverbial subordinate clause were not considered because the number with subject NP deletion found in the corpus was scarce.

a) Conditional

(3.16) *O compositor*$_i$ *Alceu Valença teria conversado com FHC*, *caso* Ø$_i$ *tivesse tido chance*

"The composer Alceu Valença had talked with FHC, if [he] had had a chance"

The non finite construction of the conditional type can be: i) with the verb in infinitive (3.17)[28]; ii) with the verb in the past participle (3.18); and iii) with the verb in gerund (3.19).

(3.17) *O óvulo*$_i$ *não poderá ser fecundado sem* Ø$_i$ *receber o devido tratamento*

"The ovum cannot be fertilized without receiving proper treatment"

(3.18) *Ele*$_i$ *conseguirá passar no exame* Ø$_i$ *estudando muito*

'He will pass in the exam studying a lot'

(3.19) *Eles*$_i$ *teriam tido outro comportamento* Ø$_i$ *reconhecido os seus direitos*

'They have had other behavior recognized their rights'

b) Causal

(3.20) *Como ela*$_i$ *a conhece bem*, Ø$_i$ *não fez nada*

"As she know her well, [she] did nothing"

The non finite construction of the causal type can be: i) with the verb in infinitive (3.21), and ii) with the verb in gerund (3.22).

(3.21) *Ele*$_i$ *não poderá voltar ao trabalho por* Ø$_i$ *estar doente*

"He cannot return to work for being sick"

(3.22) *Ele*$_i$ *dispensou-o* Ø$_i$ *desconfiando de suas palavras*

"He dropped him suspecting of his words"

c) Finality

(3.23) *As importações*$_i$ *são rigorosamente controladas para que* Ø$_i$ *não ultrapassem as exportações*

"Imports are strictly controlled in order that [they] do not exceed exports"

On the non finite construction of the finality type, the verb is in infinitive (3.24).

(3.24) *Ele*$_i$ *chegou cedo para* Ø$_i$ *ajudá-los*

"He arrived early to help them"

d) Concessive

(3.25) *Os rios*$_i$ *não secam, embora* Ø$_i$ *tenham o seu volume de água diminuído*

'The rivers do not dry, although [the rivers] has their volume of water decreased'

The non finite construction of the concessive type can be: i) with the verb in infinitive (3.26), ii) with the verb in gerund (3.27), and iii) with the verb in participle

[28] In the non finite construction of the conditional clause with the verb in the infinitive, the conjunctions *a*, *no caso* and *na condição de* can be used or not.

(3.28).

(3.26) *Ele*$_i$ *conquistou diversos prêmios apesar de* Ø$_i$ *ser jovem*

"He won several awards in spite of being young"

(3.27) *Os rios*$_i$ *não secam mesmo* Ø$_i$ *tendo o seu volume de água diminuído*

"The rivers do not dry even though the volume of water decreased"

(3.28) *Ele*$_i$ *não se entregou mesmo* Ø$_i$ *perseguido pela polícia*

"He did not surrender even chased by police"

e) Time

(3.29) *Lula*$_i$ *evitava os debates quando* Ø$_i$ *liderava as pesquisas*

"Lula avoided the debate when [he] led the poll"

The non finite construction of the time type can be: i) with the verb in infinitive (3.30), and ii) with the verb in gerund (3.31).

(3.30) *Os amigos*$_i$ *verificaram o valor do prêmio antes de* Ø$_i$ *fazerem as apostas*

Friends checked the prize before [they] make the bet

(3.31) *O João*$_i$ *viu a Maria* Ø$_i$ *olhando pela janela*

John saw Mary looking out the windows

f) Consecutive

(3.32) *A artéria*$_i$ *seria capaz de se dilatar tanto que* Ø$_i$ *até estouraria*

"The artery would be able to expand so much that [the artery] even burst"

The cases of the non finite consecutive type in which the NP subject is deleted were not considered in this dissertation because the numbers of occurrences are not many.

3.2.5 Lexically constraint coreference (control verbs)

A particular problem of anaphora resolution is presented by verbs that impose constraints on the reference of the arguments in the subordinate clause. These are called control verbs (Gross, 1975). For example:

(3.33) *O Pedro*$_i$ *queria* Ø$_i$ *ir ao cinema*

'Peter wanted to go to the movies'

(3.34) *O Pedro mandou lavar a louça*

'Peter ordered to wash the dishes'

In the sentence (3.33), the subject in the subordinate clause is obligatorily coreferent to the subject of the main verb while in sentence (3.34) they cannot be coreferent. This information must be encoded in the lexicon so that it may be used in

anaphora resolution. In section 3.5.6 we present a solution that integrates subcategorization information in zero anaphora resolution rules to deal with these cases.

3.3 Methods

We began by a systematic survey of syntactic patterns in order to identify the linguistic situations where subject NP deletion occurs and the conditions governing its deletion. Based on this survey, rules were defined and implemented in the XIP parser.

As an example, a general rule to recovery the deleted NP subject could determine that under a coordinative conjunction the zeroed NP subject on the second clause is the same NP subject of the first clause, if both have the same gender-number agreement.

(3.35) *O terremoto$_i$ matou mais de 200 pessoas e Ø$_i$ deixou milhares de pessoas desabrigadas*

'The earthquake killed more than 200 peoples and (Ø$_i$ - the earthquake) has left thousands of people homeless'

In sentence (3.35) there is the NP *o terremoto* 'the earthquake', the verb *matou* 'killed', the coordinative conjunction *e* 'and', and the verb *deixou* 'has left'. As the verbs (*matou* 'killed' and *deixou* 'has left) are in the third singular person and the NP (*o terremoto* 'the earthquake') is singular too, then the subject NP zeroed is the same subject NP of the first sentence.

Regarding the subordinate adverbial clauses, the recovery of their zeroed NP subject can be done through the same idea presented on the rule described above, since the 'antecedent' has already been explicit in the fronted clause[29].

(3.36) *O João$_i$ morou na França quando Ø$_i$ era adolescente*

'John lived in French when [he] was teenager'

(3.37) *Quando o João$_i$ era adolescente*, *Ø$_i$ morou na França*

'When John was teenager, [he] lived in French'

(3.38) *Quando Ø$_i$ era adolescente, o João$_i$ morou na França*

'When John was teenager, [he] lived in French'

[29] In the framework of Harris these reductions take flew at different stamps of the sentence concatenations process.

Thus, according to the scope of this dissertation only the sentences (3.36) and (3.37) will be considered to establish the rules; the sentence (3.38) will not be considered because the NP subject is in the second clause (cataphora).

The different order of the clauses presented in the sentences (3.36), (3.37) and (3.38) does not imply in different meaning. The difference among them is only the standpoint of pragmatics, and the emphasis is given to the NP subject written or in the first or in the second clause becoming the clause with the NP subject highlighted.

3.4 Corpus

In general, the use of corpora, among other things, serves to observe (and propose) linguistics hypothesis (in this case, formalized as rules), to optimize them and to finally evaluate them (or the approaches based on those rules) (Mitkov, et al., 1999).

To our knowledge, there is no available corpus marked up with deleted subject NPs for Portuguese[30]. Because of this lack on linguistic resources, an annotated corpus has been built for this study. The main purpose of this corpus is: the correct identification of the zero anaphor and of its antecedent (Pereira, 2009: 53).

The purpose of this linguistic resource is to promote the use of automatic discovery of linguistic parameters for anaphora resolution systems.

Two corpora were developed in order to correctly resolve zero anaphora. The corpora were provided in raw text format, but the annotation adopted can be easily converted into other formats.

3.4.1 The ZAC corpus

The Zero Anaphora Corpus (ZAC) consists on a set of full and partial texts retrieved from the web, or digitalized from books, encompassing several genres, namely journalistic and literary text from contemporary authors. This corpus was split into two parts: the training corpus with 22,385 words and the evaluation corpus with 12,827 words. Table 1 shows the breakdown per genre type of the ZAC corpus

[30] A similar corpus has been presented for Spanish (Rello and Ilisei, 2009: 209-214) but in a different theoretical framework. A corpus for anaphora resolution has been produced for Brazilian Portuguese (Collovini, *el al*, 2007: 1605-1614) but as far as we know only coreference chains between anaphors have been annotated, and no information has been made available for zero anaphors.

current content. In this table, there are the different genres texts – special report, news, chronicle, short stories and novel discriminated in the *Text Types* column; the number of the words that compound each genre – shown in the *Words* column; and the percentage corresponding to the total number of words that each gender has regarding to the total number of words in the corpus.

Table 1: Content of the ZAC corpus

	Training corpus		Evaluation corpus		ZAC corpus	
Text Types	Words	%	Words	%	Words	%
Special Report	10,272	46%	5,519	43%	15,791	45%
News	905	4%	864	7%	1,769	5%
Chronicle	5,416	24%	2,969	23%	8,385	24%
Fiction (short story)	2,029	9%	1,198	9%	3,227	9%
Fiction (novel)	3,763	17%	2,277	18%	6,040	17%
Total	22,385		12,827		35,212	

The corpus was manually annotated. The evaluation corpus was annotated separately and was only used for testing. General notation is as follows: zero anaphors are marked by a zero symbol '0' inside brackets [ ], followed by an equal sign '=' and the arrow symbols '<' and '>', corresponding to anaphora (3.39) and cataphora (3.40) relations, respectively, and a word indicating the head of the antecedent noun phrase (NP).

(3.39) *Um forte terremoto (6 graus na escala Richter) sacudiu ontem Taiwan*, **[0=< terremoto]** *provocando uma morte e ferimentos em duas pessoas*

'A strong earthquake (measuring 6 degrees on the Richter scale) shook Taiwan yesterday; the earthquake caused one death and injured two people'

(3.40) *Ao* **[0=>descobertas]** *apontarem para a cura de doenças atacando-as na escala infinitesimal dos genes, as novas descobertas da ciência representam um novo marco na linha de pensamento iniciada no século XIX pelo naturalista inglês Charles Darwin, autor da teoria da evolução*

'Pointing to the cure of diseases by attacking them in the infinitesimal scale of genes, the new discoveries of science represent a new milestone in the line of thought that has been started in the nineteenth century by the English naturalist Charles Darwin, author of the theory of evolution'

The criteria used to annotate the corpus are given in Appendix 2.

Preliminary results

Preliminary results from the annotation process of the ZAC corpus are presented in the tables below. In these tables, besides the columns *Text Types* and *Words* (described above), there are the column *Total marks* - with the total number of tags annotated in each genre text, and the columns *indef*, *impers*, *1p* and *3p* with the total number of cases found in the corpus.

The column *indef* represents cases in which the subject is an indefinite zeroed subject (3.41) and the tag used is **[0=indef]**:

(3.41) *Apesar de todos os avanços na ciência da genética, apenas dentro de uma ou duas décadas será possível* **[0=indef]** *prevenir o aparecimento de doenças [0=indef] auscultando os genes, ou* **[0=indef]** *produzir remédios personalizados que ajam sobre o genoma específico de um paciente*

'Despite all the advances in genetic science, only in one or two decades will it be possible to prevent diseases from appearance by checking the genes, or to produce personalized medicines acting on the specific genome of a patient'

In the column *impers*, it is the cases in which the verb is impersonal (3.42) and, therefore, there is not subject. These cases were marked up with the tag **[0=impers]**.

(3.42) "**[0=impers]** *Há uma perigosa tendência a* [0=indef] *fazer correlações entre etnia, crime e predisposição genética", alerta Pamela Sankar, professora de bioética da Universidade da Pensilvânia.*

'"There is a dangerous tendency to establish correlations between ethnic origin, crime and genetic predisposition", alerts Pamela Sankar, Bioethics professor at Pennsylvania University.'

The column *1p* shows the cases in which the subject is also an indefinite but there is a systematic ambiguity with first person *nós* 'we' (3.43). These cases were annotated using the tag **[0=1p]**. The difference between *1p* cases and *indef* cases is that while in the first case the verb is in the first person plural (3.43), in the second case the verb is on the bare infinitive (3.41).

(3.43) *As descobertas são impressionantes.* **[0=1p]** *Conseguimos informações preciosas sobre os genes, as marcas epigenéticas e as mudanças do genoma ao longo da vida, o que dá início a uma revolução*

'The findings are impressive. We got valuable information about the genes, the epigenetic markings and the changes of the genome throughout life, which initiates a revolution'

Finally, in the column *3p*, we find the cases in which the verb is in the third person plural (3.44) and the subject is interpreted as an indefinite. These cases were annotated using the tag **[0=3p]**.

(3.44) *Estou esperando o que me* **[0=3p]** *garantiram [...]*

'[I] am waiting what [they] assured me'

These columns represent the subjects that do not correspond to zero anaphors. Their identification constitutes a linguistic challenge for any anaphora resolution system. Overall, they represent 401 (26.93%)[31] from all zeroed subjects in the ZAC corpus.

Table 2: Indefinite/impersonal subjects per genre in the training corpus

Training corpus						
Text Types	Words	Total marks	indef	impers	1p	3p
Special Report	10,272	357	67	26	34	0
News	905	21	3	1	0	0
Chronicle	5,416	243	36	11	37	6
Fiction (short story)	2,029	99	2	3	4	5
Fiction (novel)	3,763	210	4	18	8	17
Total	22,385	930	112	59	83	28

Table 3: Indefinite/impersonal subjects per genre in the evaluation corpus

Evaluation corpus						
Text Types	words	Total marks	indef	impers	1p	3p
Special Report	5,519	181	14	16	7	3
News	864	31	5	3	0	0
Chronicle	2,969	152	5	6	6	2
Fiction (short story)	1,198	47	2	8	1	11
Fiction (novel)	2,277	148	3	8	11	8
Total	12,827	559	29	41	25	24

Table 4: Indefinite/impersonal subjects per genre in the ZAC corpus

ZAC corpus (Training + Evaluation)						
Text Types	words	Total marks	indef	impers	1p	3p
Special Report	15,791	538	81	42	41	3
News	1,769	52	8	4	0	0
Chronicle	8,385	395	41	17	43	8
Fiction (short story)	3,227	146	4	11	5	16
Fiction (novel)	6,040	358	7	26	19	25
Total	35,212	1,489	141	100	108	52

[31] In the training corpus they represent 30,32% (282) and in the evaluation corpus they represent 21,29% (119).

In Table 2, the indefinite subjects (*[0=indef]*) correspond to 12.04% of all marks; the impersonal subjects (*[0=impers]*) correspond to 6.34% of all marks; and the *1p* and *3p* tags correspond to 8.92% and 3.01% of all marks respectively.

Table 3 follow the same idea presented above and the results were: 5.18% for *indef*, 7.33% for *impers*, 4.47% for *1p*, and 4.29% for *3p*.

Finally, Table 4 presents the results for the entire corpus (ZAC corpus): 9.46% for *indef*, 6.71% for *impers*, 7.25% for *1p* and 3.49% for *3p*.

The *1p* (first person plural) and *3p* (third person plural) indefinite zeroed subject types may be targeted by using the verbal inflection as a clue in cases in which any other candidate antecedent NP is absent. They represent around 11.93% in the training corpus, 8.76% in the evaluation corpus and 10.74% in the whole corpus.

Indefinite zeroed subjects, without *1p* or *3p* inflection associated, are harder to identify and in this cases the verb is usually in the bare infinitive.

Finally, the identification of impersonal constructions heavily relies on the resolution of other syntactic issues such as auxiliary constructions and temporal expressions.

Table 5, Table 6 and Table 7 presents the breakdown of anaphoric and cataphoric zero anaphora per genre in the training corpus, in the evaluation corpus and in the entire corpus respectively. In the tables there are also the distinction of the anaphora and the cataphora with intra- (< , >) and intersentencial (<< , >>) antecedent.

Table 5: Anaphora/cataphora breakdown per genre in the training corpus

Training corpus				
Text Types	<	<<	>	>>
Special Report	170	51	14	0
News	15	1	1	0
Chronicle	83	61	1	1
Fiction (short story)	31	49	3	0
Fiction (novel)	86	67	7	0
Subtotal	385	229	26	1
Total	614		27	

Table 6: Anaphora/cataphora breakdown per genre in the evaluation corpus

Evaluation corpus				
Text Types	<	<<	>	>>
Special Report	105	23	6	0
News	19	1	3	0
Chronicle	73	54	4	1
Fiction (short story)	13	16	1	0
Fiction (novel)	85	32	1	0
Subtotal	295	126	15	1
Total	421		16	

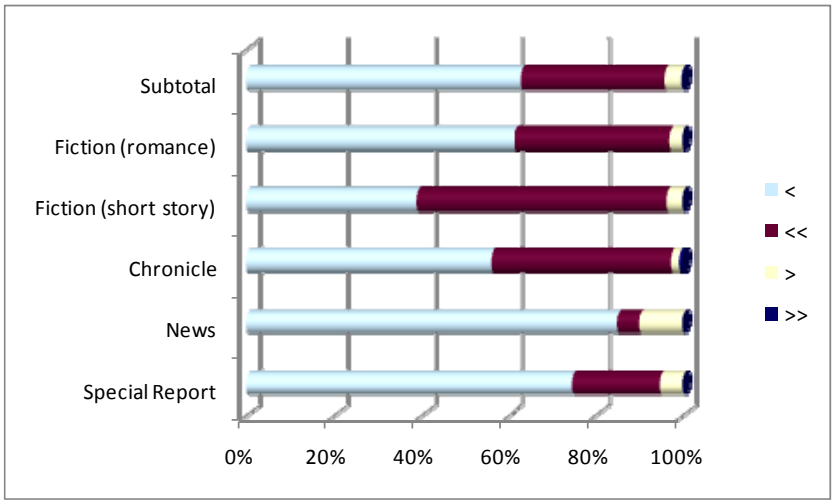
Table 7: Anaphora/cataphora breakdown per genre in the ZAC corpus

ZAC corpus (Training + Evaluation)				
Text Types	<	<<	>	>>
Special Report	275	74	20	0
News	34	2	4	0
Chronicle	156	115	5	2
Fiction (short story)	44	65	4	0
Fiction (novel)	171	99	8	0
Subtotal	680	355	41	2
Total	1,035		43	

As one can see, cataphora (>, >>) is a relatively rare phenomenon, affecting a little over 3% of all anaphors in the corpus (2.90% in the training corpus and 2.86% in the evaluation corpus). Intrasentencial anaphora (<) represents 45.67% (41.40% in the training corpus and 52.77% in the evaluation corpus) while intersentencial anaphora (<<) constitutes 23.84% (24.62% in the training corpus and 22.54% in the evaluation corpus).

There seems to be little difference among genres as far as anaphora/cataphora ratio is concerned. On the other hand distinction between intra- and intersentencial anaphors is much clearer as one can see from Figure 3. News and special reports genres show clear predominance of intrasentencial anaphora (around 80 and 70%, respectively); fiction (novel) and chronicle show average intrasentencial anaphora (around 60 and 50%, respectively); and finally fiction (short stories) only presents 40% intrasentencial anaphora. However, since the corpus is relatively small and only includes a few genres these differences may vary if a larger corpus was available and if it included other genre types.

Figure 3: Anaphora/cataphora breakdown per genre in the ZAC corpus



The 23 special cases of `0(clause)` (7 cases) and `0(que)` (15 cases) represent a very rare phenomenon (1.5% of all zero subjects). The last resort '?' notation for 39 cases, where it was impossible to arrive at a positive identification of antecedent NP represents 2.6%.

3.4.2 The Sentence corpus

Another corpus – the Sentence corpus – consist on a set of sentences retrieved from the CETENFolha[32] (***C**orpus de **E**xtractos de **T**extos **E**lectrónicos **N**ILC/**Folha** de S. Paulo* 'Corpus of electronic texts extracts from the daily Brazilian Portuguese newspaper *Folha de São Paulo*' compiled by NILC – Interinstitutional Center for Computational Linguistic) (Pinheiro and Aluísio, 2003)). In addition, another set of sentences[33] was specially constructed in order to test the rule of control verbs and the rule of attributes[34]. This set is formed of sentences that show diverse possibilities of word order and involve diverse syntactic structures with different words which belong to the same verb or adjective categories.

This corpus was provided in order to select different cases of zero anaphora that did not appear on the ZAC corpus. This corpus was annotated with the same symbols used on the ZAC corpus. The total number of the zero anaphora annotated on the sentence corpus was 256 cases.

[32] Available at: http://www.linguateca.pt/
[33] This set can be seen in Appendix 3.
[34] To see the rules for control verbs and attributes, see 3.5 section.

3.5 Linguistically motivated rules

After a systematic linguistic analysis of the zero anaphora cases presented above, some general rules were defined and implemented[35]. These rules will be briefly presented here[36]. Before that, it should be noted that these AR rules rely on the rule-based grammar for Portuguese (Mamede et al., 2010) that has been developed so far at L2F/INESC ID Lisboa and implemented in the XIP parser. In fact, much work had already been done in the NLP processing chain, in particular on the chunking module and on the set of dependencies that AR rules use. On the other hand, exploring this new venue on the grammars' development showed some previously undetected problems that we took the opportunity to help solving.

3.5.1 Coordinate clause

Coordination is one of the most important contexts for anaphoric reduction. However, parsing coordination is a very challenging task because of long range constraints, different syntactic levels involved, and the different repetition constraints on the two members of a coordinative operator (Harris, 1991). Besides, coordination can also involve certain phenomena, such as apposition, not often included in this part of grammar.

Consider the following sentence:

(3.45) *Essas células*$_i$ *viajariam pelo corpo até os órgãos sexuais e de lá Ø*$_i$ *passariam às gerações seguintes*

'These cells would travel through the body to the sexual organs and from there [they] would pass on to the subsequent generations'

The rule that deals with coordinate sentences like the example above is shown in Figure 4:

[35] Linguistically motivated rules were implemented by Prof. Nuno Mamede in the computational grammar developed for the Portuguese language at L2F/INESC ID Lisboa using the XIP parser. I would like to acknowledge him for his help and patience in this interactive process of bridging linguistic, often theoretical, concepts to the parser formalization.

[36] These rules can also be seen in Appendix 4.

Figure 4: Rule for the coordinate clause

```
|     #1[verb],     ?*,     CONJ[coord];PUNCT[lemma:";"];PUNCT[lemma:":"],
?*[verb:~,sc:~], #3[verb] |

   if  (  HEAD(#4,#1)  &  VDOMAIN(?,#4)  &  SUBJ(#4,#5)  &  HEAD(#6,#3)  &
VDOMAIN(#6,#7) & ~SUBJ(#7,?) &

     ((#5[person]:#7[person] & #5[number]:#7[number]) || (#5[person:~] &
#7[3p] & #5[number]:#7[number] & ~COORD(?,#5)) || (#5[person:~] & #7[3p,pl]
& COORD(?,#5)) || #7[person:~])

   )

    SUBJ[pre=+,anaph0=+](#7,#5)
```

In short, this rule states that in a sentence with two coordinate clauses, if the verb of the first clause has an explicit subject and the verb of the second clause has not, then creates a zero-anaphoric subject dependency, and consider that the subject of the first verb is coreferent of the subject of the second verb.

The output of the sentence is shown in the Figure 5:

Figure 5: Output of the coordinate rule (sentence (3.45))

```
MAIN(viajariam)                          MOD_POST(viajariam,corpo)
VDOMAIN(viajariam,viajariam)             MOD_POST(viajariam,órgãos)
VDOMAIN(passariam,passariam)             MOD_POST(viajariam,lá)
MOD_POST(órgãos,sexuais)                 MOD_POST(passariam,gerações)
MOD_POST(gerações,seguintes)             SUBJ_PRE(viajariam,células)
MOD_POST(corpo,órgãos)                   SUBJ_PRE_ANAPH0(passariam,células)

8>TOP{NP{Essas células} VF{viajariam} PP{por o corpo} PP{até os órgãos}
AP{sexuais} e PP{de lá} VF{passariam} PP{a as gerações} AP{seguintes} .}
```

Among the dependencies extracted by the system, we find the unary MAIN dependency (usually the first finite verb); the binary dependencies: DETD, between the determiner and the NP head noun; PREPD, between the preposition and the PP head; VDOMAIN between the first verb and the last verb of a verbal chain, i.e. a sequence of auxiliaries and the main verb (Baptista et al. 2010); MOD, between the head of previous chunk and the head of any kind of complements or adjuncts that may be attached to this chunk[37]; SUBJ, between a verb and its surface subject; and the features _PRE and _POST that function as dependency features indicating that the chunk is on the left (PRE) or on the right (POST) of its governor in the sentence. Besides, there is the feature _ANAPH0 which indicates that a zero anaphor has been reconstructed. This feature is used with the SUBJ dependency.

[37] At this stage only limited linguistic information on subcategorization is being used. At a latter stage, the espurious MOD dependencies are trimmed out.

Notice that the rule invokes previously defined dependencies such as VDOMAIN and SUBJ, that is, it does not aim individual verbs but eventually any verbal chain and it requires that the subject of the second verb be identified before. Notice also that the fifth line of the rule imposes person-number agreement between the subject of the first verb and the second. This agreement constraint is also used in other rules.

Besides this rule, several existing rules already dealt with local coordination. In these rules, the _ANAPH0 feature was added. However, for coordinate NPs it was necessary to extend this feature in order to be able to capture the following coordinated clause.

(3.46) *O João e a Maria*$_i$ *comeram o bolo mas Ø*$_i$ *ficaram com fome*

'John and Mary ate the cake but [they] were hungry'

This is done by the following two rules:

Figure 6: Rule for coordinate NPs

```
if ( SUBJ[anaph0](#2,#1) && coord(#3,#1) & coord(#3,#4) && ~SUBJ(#2,#4) )
    SUBJ[anaph0=+,pre=+](#2,#4)
```

Figure 7: Rule for coordinate NPs

```
if ( SUBJ[anaph0](#2,#1) && coord(#3,#1) & coord(#3,#4) &&
^SUBJ[anaph0:~](#2,#4) )
    SUBJ[anaph0=+](#2,#4)
```

These rules state, in short, that if two coordinate NP are identified as the subject of the first verb in a coordinate clause, and if the first NP is already considered the antecedent of the subject of the verb in the second coordinate clause, then both NPs are anaphoric subjects of the second verb. This happens because coordination is dealt with by two dependencies, linking each NP to the coordinative conjunction so that each one of those NPs are related to its verb by a separate SUBJ dependency and the _ANAPH0 feature also needs to be duplicated.

3.5.2 Subordinate clause

Subordinate adverbial clauses are also a major factor for subject NP deletion. Besides, the number of subordinate conjunctions is larger than coordinate conjunctions, so the matter of lexical coverage becomes an important aspect for any rule-based AR system.

The second general rule (Figure 8) deals with subordinate clauses. The main difference between the second and the first rule is related to the conjunction; while in

the first rule there is a coordinate conjunction (CONJ[coord]), in this second rule, a subordinate clause is indicated by the SC (subclause) chunk. This chunk is construed *grosso modo* by linking a subordinate conjunction to the first finite verb to its right.

Figure 8: Rule for the subordinate clause

```
| #1[verb], ?*[verb:~], SC{?*, ?#3[verb,last]} |

   if  (  HEAD(#4[s_qufconj:~],#1)  &  VDOMAIN(?,#4)  &  SUBJ(#4,#5)  &
HEAD(#6,#3) & VDOMAIN(#6,#7) & ~SUBJ(#7,?)  &

     ((#5[person]:#7[person] & #5[number]:#7[number]) || (#5[person:~] &
#7[3p] & #5[number]:#7[number] & ~COORD(?,#5)) || (#5[person:~] & #7[3p,pl]
& COORD(?,#5)) || #7[person:~])

   )

    SUBJ[pre=+,anaph0=+](#7,#5)
```

The rule can be described as follows: in a subordinate sentence, if the verb of the main clause has an explicit subject and the verb of the secondary (subordinate) clause has not, a zero-anaphoric subject dependency is created and the subject is reconstituted from the subject of the main clause. Therefore, this rule is activated only after the module that deals with the identification of the SC chunk.

Consider the following sentence:

(3.47) *O senhor*$_{i}$ *estava fingindo que esperava um ônibus, em atitude suspeita, quando Ø*$_{i}$ *suspeitou destes dois agentes da lei ao seu lado*

'The old man was pretending to be waiting for a bus, in suspicious manner, when [the old man] suspected the two law enforcement officials beside him'

The output of the sentence is shown in the Figure 9:

Figure 9: Output of the subordinate rule (sentence (3.47))

```
MAIN(fingindo)                      MOD_POST(suspeitou,lado)
VDOMAIN(esperava,esperava)          POSS_PRE(lado,seu)
VDOMAIN(suspeita,suspeita)          SUBJ_PRE(fingindo,senhor)
VDOMAIN(suspeitou,suspeitou)        SUBJ_PRE_ANAPH0(esperava,senhor)
VDOMAIN(estava,fingindo)            SUBJ_PRE_ANAPH0(suspeita,senhor)
MOD_POST(agentes,lei)               SUBJ_PRE_ANAPH0(suspeitou,senhor)
MOD_POST(agentes,lado)              CDIR_POST(esperava,ônibus)
MOD_POST(lei,lado)                  CDIR_SENTENTIAL_POST(fingindo,esperava)
MOD_POST(fingindo,atitude)          SUBORD_COMPLETIV(que,esperava)
MOD_POST(esperava,atitude)          SUBORD(quando,suspeitou)
MOD_POST(suspeitou,agentes)         EMBED(fingindo,esperava)
MOD_POST(suspeitou,lei)             INTROD_COMPLETIV(fingindo,que)

151>TOP{NP{O senhor} VASP{estava} VGER{fingindo} SC{que VF{esperava}} NP{um
ônibus} , PP{em atitude} VF{suspeita} , SC{quando VF{suspeitou}} PP{de
estes dois agentes} PP{de a lei} PP{a o seu lado} .}
```

Besides the dependencies that have already been explained above, in the output of the sentence (3.47), there are also the following dependencies: CDIR,

which links a verb and its direct object; SUBORD, which links the beginning of a subordinate phrase to the first verb of that subordinate phrase; EMBED, which links the main verb of an embedded clause to its governor; and INTROD, which links a verb to the conjunction that starts the embedded subclause. There are also the features: _SENTENTIAL, which is added to SUBJ, CDIR and MOD dependencies in order to indicate that the subject, direct object or modifier is a subclause; and _COMPLETIV, which is used to mark the fact that an embedded clause is a completive (nominal subclause).

Notice that the parser incorrectly assigned the dependency SUBJ_PRE_ANAPH0(suspeita,senhor) because the word *suspeita* 'suspicious' is ambiguous and has been classified as a verb by the POS tagger. It should have been tagged as an adjective and because of that the chunker produced a VF chunk instead of an AP.

3.5.3 Anteposition of the subordinate clause

In general, subordinate clauses can be moved to the front of the main clause:

(3.48) *Quando alguém*$_i$ *começa a incomodar, Ø*$_i$ *é ignorado ou deletado*

'When someone begins to bother you, is ignored or deleted'

In the sentence (3.48), the subordinate clause *Quando alguém começa a incomodar* 'When someone begins to bother you' has been fronted to the beginning of the main clause "*é ignorado (…)*" 'is ignored'. The subject of the main clause has been zeroed since it has already appeared.

This transformation requires a new rule to capture the zeroed subject (Figure 10).

Figure 10: Rule for the anteposition of the subordinate clause

```
|    ?*[verb],    SC{?*,    ?#1[verb,last]},    ?*[sc:~],    PUNCT[comma],
?*[verb:~,sc:~], ?#3[verb] |

   if  (  HEAD(#2,#1)  &  SUBORD(?,#2)  &  VDOMAIN(#2,#4)  &  SUBJ(#4,#5)  &
HEAD(#6,#3) & VDOMAIN(#6,#7) & ~SUBJ(#7,?) )

    SUBJ[pre=+,anaph0=+](#7,#5)
```

The rule takes into account the following situation: if the sentence begins with a subordinate conjunction and the verb of this subordinate clause has an explicit subject; and if the verb of the main clause has no subject dependency yet; if the two

clauses are separated by comma ‘,’[38] ; then a zero-anaphoric subject dependency is created and the subject is reconstituted from the subject of the first clause.

The output of sentence (3.48) is presented in the Figure 11.

Figure 11: Output of the anteposition rule (sentence (3.48))

```
MAIN(ignorado)                          SUBJ_PRE(incomodar,alguém)
VDOMAIN(começa,incomodar)               SUBJ_PRE_ANAPH0(ignorado,alguém)
VDOMAIN(é,ignorado)                     SUBORD(Quando,começa)

51>TOP{SC{Quando NP{alguém} VASP{começa a}} VINF{incomodar} , VCOP{é}
VCPART{ignorado} ou NP{deletado} .}
```

As we can see, the dependency SUBJ_PRE_ANAPH0(ignorado,alguém) was assigned correctly to the verb chain *é ignorado* ‘is ignored’[39]. However, as the rule invokes previously defined dependencies (section 3.5.1), the same syntactic structure presented in the sentence above can fail. Consider the following sentence:

(3.49) *Quando Raul*$_i$ *também fechou sua janela, Ø*$_i$ *encontrou Luiz abrindo a caixa da luneta, e Marina acalmando Thaíssa, ambas sentadas na cama de Raul*

‘When Raul also closed his window, [he] found Luiz opening the box of the telescope, and Marina appeasing Thaíssa, both sitting on the Raul’s bed’

and the output of the sentence (3.48) is shown in the Figure 12:

Figure 12: Output of the anteposition rule (sentence (3.49))

```
MAIN(encontrou)                                 POSS_PRE(janela,sua)
VDOMAIN(fechou,fechou)                          SUBJ_PRE(fechou,Raul)
VDOMAIN(encontrou,encontrou)                    SUBJ_PRE(encontrou,janela)
VDOMAIN(abrindo,abrindo)                        SUBJ_PRE(abrindo,Luiz)
VDOMAIN(acalmando,acalmando)                    SUBJ_PRE_ANAPH0(acalmando,janela)
MOD_PRE(fechou,também)                          CDIR_POST(abrindo,caixa)
MOD_POST(ambas,sentadas)                        CDIR_POST(acalmando,Thaíssa)
MOD_POST(caixa,luneta)                          CDIR_POST(fechou,janela)
MOD_POST(ambas,Raul)                            CDIR_POST_INF(encontrou,abrindo)
MOD_POST(sentadas,cama)                         SUBORD(Quando,fechou)
MOD_POST(sentadas,Raul)                         NE_INDIVIDUAL_PEOPLE(Raul)
MOD_POST(cama,Raul)                             NE_INDIVIDUAL_PEOPLE(Luiz)
MOD_SENTENTIAL_POST_GERUND(encontrou,abrindo)   NE_INDIVIDUAL_PEOPLE(Marina)
MOD_SENTENTIAL_POST_GERUND(abrindo,acalmando)   NE_INDIVIDUAL_PEOPLE(Raul)

145>TOP{SC{Quando NP{Raul} ADVP{também} VF{fechou}} NP{sua janela} ,
VF{encontrou} NP{Luiz} VGER{abrindo} NP{a caixa} PP{de a luneta} , e
NP{Marina} VGER{acalmando} NP{Thaíssa} , NP{ambas} AP{sentadas} PP{em a
cama} PP{de Raul} .}
```

Notice that the parser incorrectly assigned the SUBJ_PRE(encontrou,janela) dependency where it should have assigned the

[38] The requirement of comma was meant to limit the scope of the rule.
[39] The past parciple *deletado* ‘deleted’ was incorrectly tagged as a noun. This word is only usual in BP.

dependency _ANAPH0 between the verb *encontrou* ‘found’ and the NP *Raul* [40]. This happened because at this stage the control verb *encontrar* ‘to find’ was not yet encoded in the lexicon (see section 3.5.6).

Still regarding to the anteposition structure, there are also cases in which the subject of the fronted subordinate clause may also be zeroed and the subject of the main clause be kept, like in sentence (3.50) and (3.51), which may be considered a case of cataphora[41]:

(3.50) *Depois de Ø$_i$ cair pela metade entre 2000 e 2006, o desmatamento$_i$ voltou a crescer no verão amazônico que se encerrou em outubro – 14 mil km² de florestas foram abaixo no último ano, o que dá quase um Líbano e meio*

‘After falling by half between 2000 and 2006, the deforestation has risen again on the Amazonic summer that ended on October – 14 thousand Km^2 went down last year, which is about one Lebanon and a half’

(3.51) *Apesar de Ø$_i$ superar o nível de outubro, porém, o índice$_i$ de demanda por crédito ainda é 1,2% menor que o de junho de 2008*

‘In spite of surpassing the level of October, however, the credit demand index is still 1.2% less than the one of June 2008.’

As we said above, we do not concern with cataphora but, since the rule that deals with these cases is very similar to the rule shown in Figure 13 we took the opportunity to formalized it as well:

Figure 13: Rule for the anteposition of the subordinate clause (cataphora)

```
|    ?*[verb],    SC{?*,    ?#1[verb,last]},    ?*[sc:~],    PUNCT[comma],
?*[verb:~,sc:~], ?#3[verb] |

   if  ( HEAD(#2,#1)  &  SUBORD(?,#2)  &  VDOMAIN(#2,#4)  &  ~SUBJ(#4,?)  &
HEAD(#6,#3) & VDOMAIN(#6,#7) & SUBJ(#7,#5) )

    SUBJ[post=+,anaph0=+](#4,#5)
```

To put it simply this rule states that after a subclause chunk SC and a string ending in comma; and if this subclause does not have a subject dependency extracted yet; and if after the comma there is no other subclause before a given verb; then take the subject of the latter and create a subject dependency between this one and the verb of the initial subclause. Notice that the comma has been enforced in

[40] In this example, the unary dependency NE is also extracted for the Named Entity, which is further classified, using the Portuguese NER ontology adopted for HAREM2 (Mota and Santos, 2008).
[41] In the framework of Harris (1991), the zeroing of the subject of the subordinate clause takes place before its fronting, therefore in this perspective it is not exactly a case of cataphora.

order to define a rightmost limit to the fronted subclause. Notice also the new dependency SUBORD already built over the SC chunk[42].

The output of sentences (3.50) and (3.51) is showed below:

Figure 14: Output of the anteposition rule (cataphora) (sentence (3.50))

```
MAIN(foram)                          MOD_POST(foram,abaixo)
VDOMAIN(cair,cair)                   SUBJ_PRE(crescer,desmatamento)
VDOMAIN(encerrou,encerrou)           SUBJ_SENTENTIAL_PRE_INF(foram,crescer)
VDOMAIN(foram,foram)                 SUBJ_POST_ANAPH0(cair,desmatamento)
VDOMAIN(dá,dá)                       CDIR_POST(dá,Líbano)
VDOMAIN(voltou,crescer)              CDIR_POST(foram,o)
MOD_PRE(crescer,pela metade)         SUBORD(que,encerrou)
MOD_POST(verão,amazônico)            SUBORD(que,dá)
MOD_POST(crescer,verão)              SUBORD(Depois de,cair)
MOD_POST(crescer,outubro)            NE_LOCAL_COUNTRY_ADMIN_AREA(Líbano)
MOD_POST(crescer,florestas)          NE_TEMPO_INTERVAL(entre 2000 e 2006)
MOD_POST(foram,último ano)           NE_T-ABSOLUT_TEMPO_DATE(em outubro)
MOD_POST(cair,2006)                  NE_TREF-ENUNC_TEMPO_DATE(em o verão)
MOD_POST(encerrou,outubro)           NE_TREF-TEXT_TEMPO_DATE(em o último ano)
MOD_POST(encerrou,florestas)         NE_QUANT_NUM(14 mil km)

1>TOP{SC{Depois de VINF{cair}} ADVP{pela metade} PP{entre 2000 e 2006} ,
NP{o desmatamento} VASP{voltou a} VINF{crescer} PP{em o verão}
AP{amazônico} SC{que NP{se} VF{encerrou}} PP{em outubro} - NP{14 mil km} ²
PP{de florestas} VF{foram} ADVP{abaixo} PP{em o NOUN{último ano}} , NP{o}
SC{que VF{dá}} NP{quase um Líbano} e AP{meio} .}
```

Figure 15: Output of the anteposition rule (cataphora) (sentence (3.51))

```
MAIN(%)                              SUBJ_PRE(é,índice)
PREDSUBJ(é,%)                        SUBJ_POST_ANAPH0(superar,índice)
VDOMAIN(superar,superar)             CDIR_POST(superar,nível)
VDOMAIN(é,é)                         SUBORD(Apesar de,superar)
MOD_PRE(é,ainda)                     ATTRIB(índice,%)
MOD_POST(nível,outubro)              NE_T-ABSOLUT_TEMPO_DATE(de outubro)
MOD_POST(índice,demanda)             NE_T-ABSOLUT_TEMPO_DATE(de junho de 2008)
MOD_POST(índice,crédito)             NE_QUANT_NUM(1,2 %)
MOD_POST(demanda,crédito)

2>TOP{SC{Apesar de VINF{superar}} NP{o nível} PP{de outubro} , porém , NP{o
índice} PP{de demanda} PP{por crédito} ADVP{ainda} VF{é} NP{1,2 %}
AP{menor} que o PP{de NOUN{junho de 2008}} .}
```

3.5.4 Infinitive adverbial subordinate clause

One of the most common cases of zeroed subject anaphor happens in infinitive[43] adverbial subordinate clauses.

[42] Notice that in the output of sentences (3.50) and (3.51), the unary dependency NE is also extracted for the time expressions, places and quantity. This module was developd using XIP as parser and NE extractor (Hagège et al., 2010).

[43] Portuguese presents two infinitives: *bare* (or *impersonal*, or *non-inflected*) infinitive: *lavar* 'wash', and the *personal* (or *inflected*) infinitive: *lavar*_1st/3rdsg, *lavares*_2ndsg, *lavarmos*_1stpl, *lavardes*_2ndpl

(3.52) *Já os homens*$_i$ *se especializaram em Ø*$_i$ *estabelecer um número maior de relações, mas com um grau de intimidade menor*

'Already the men specialize in establishing a larger number of relationships, but with a lesser degree of intimacy'

To solve these cases, the following rule has been developed:

Figure 16: Rule for the infinitive adverbial subordinate clause

```
if ( MOD[post,inf,sentential](#1,#7) & SUBJ[pre](#1,#5) & ~SUBJ(#7,?) )
 SUBJ[pre=+,anaph0=+](#7,#5)
```

This rule is based on previously calculated `MOD` dependency. At this stage of the grammars, only subject and direct object argument dependencies have been created since the parser usually does not use subcategorization information associated to predicates (see section 3.2.5 for some of the first tentative in using this syntactic-semantic information). Therefore all complements that have not yet received any argumental status are treated as modifiers of the main verb.

Infinitives are captured by one of the `MOD` dependency rules, which also add the `post_inf_sentential` feature. After this, the AR rule above is straightforward: if this particular kind of `MOD` does not have a subject and if there is any subject dependency with subject-verb normal order in some previous moment in the sentence then take this subject and create the subject dependency of the infinitive adverbial subordinate clause.

The output of the sentence (3.52) is shown in Figure 17:

Figure 17: Output of the infinitive adverbial rule (sentence (3.52))

```
MAIN(especializaram)                    MOD_POST(grau,intimidade)
VDOMAIN(especializaram,especializaram)  MOD_POST(estabelecer,relações)
VDOMAIN(estabelecer,estabelecer)        MOD_SENTENTIAL_POST_INF
MOD_PRE(especializaram,Já)              (especializaram,estabelecer)
MOD_POST(número,maior)                  SUBJ_PRE(especializaram,homens)
MOD_POST(intimidade,menor)              SUBJ_PRE_ANAPH0(estabelecer,homens)
MOD_POST(número,relações)               CDIR_POST(estabelecer,número)
MOD_POST(maior,relações)

43>TOP{ADVP{Já}  NP{os  homens}  NP{se}  VF{especializaram}  VINF{em
estabelecer} NP{um número} AP{maior} PP{de relações} , mas PP{com um grau}
PP{de intimidade} AP{menor} .}
```

Notice, however, that many prepositions can also function as subordinate conjunctions as it is the case of *para* 'to' (3.53) but also *por* 'by, because' and *sem* 'without', for example. A particular case is the contraction *ao* 'to_the_m_sg' (3.54)

---

and *lavarem*_3rdpl. For the purpose of this dissertation, agreement rules on infinitives were not taken into account.

and (3.55):

(3.53) *Muitos testes*$_i$ *desse tipo servem apenas para Ø*$_i$ *criar uma neurose em torno da genética*

'Many tests of this kind serve to create a neurosis concerning genetics'

(3.54) *Agora responda: o que você*$_i$ *faria ao Ø*$_i$ *perceber que em a sua cabeça existe uma idéia que pode abalar as crenças mais profundas de quase toda a humanidade*

'Answer now: what would you do if you perceived that you have in mind an idea that could shake the most important beliefs of almost the interi human kind'

(3.55) *Ao Ø*$_i$ *apontarem para a cura de doenças, as novas descobertas*$_i$ *da ciência representam um novo marco na linha de pensamento iniciada no século XIX pelo naturalista inglês Charles Darwin, autor da teoria da evolução.*

'By showing the cure for diseases, the new discoveries of science represent a new milestone in the trend of though begun in the 19th century by the English naturalist Charles Darwin, author of the theory of the evolution.'

As it would not be wise to systematically double the POS tag for these words (both as preposition and a conjunction), some disambiguation rules were also created to produce the correct part-of-speech tag for these forms. Once this rule is in place, the output of the system for sentences (3.53), (3.54) and (3.55) is:

Figure 18: Output of the infinitive adverbial rule (sentence (3.53))

```
MAIN(servem)                          MOD_POST(criar,genética)
VDOMAIN(servem,servem)                MOD_POST(servem,apenas)
VDOMAIN(criar,criar)                  SUBJ_PRE(servem,testes)
MOD_POST(testes,tipo)                 SUBJ_PRE_ANAPH0(criar,testes)
MOD_POST(neurose,genética)            CDIR_POST(criar,neurose)
MOD_POST(servem,genética)             SUBORD_FINAL(para,criar)

17>TOP{" NP{Muitos testes} PP{de esse tipo} VF{servem} ADVP{apenas} SC{para
VINF{criar}} NP{uma neurose} PP{em torno de a genética} .}
```

Figure 19: Output of the infinitive adverbial rule (sentence (3.54))

```
MAIN(responda)                        POSS_PRE(cabeça,sua)
VDOMAIN(responda,responda)            SUBJ_PRE(faria,você)
VDOMAIN(faria,faria)                  SUBJ_POST(existe,idéia)
VDOMAIN(perceber,perceber)            SUBJ_POST(abalar,crenças)
VDOMAIN(existe,existe)                SUBJ_PRE_ANAPH0(perceber,você)
VDOMAIN(pode,abalar)                  CDIR_SENTENTIAL_POST_INF(responda,abalar)
MOD_PRE(responda,Agora)               CDIR_PRE(abalar,que)
MOD_POST(crenças,profundas)           SUBORD_TEMPORAL(ao,perceber)
MOD_POST(crenças,quase)               SUBORD(que,faria)
MOD_POST(profundas,quase)             SUBORD(que,existe)
MOD_POST(abalar,quase)                INTROD_SUPERLATIVO(crenças,mais)
```

```
92>TOP{ADVP{Agora} VF{responda} : NP{o} SC{que NP{você} VF{faria}} SC{ao
VINF{perceber}} SC{que PP{em a sua cabeça} VF{existe}} NP{uma idéia}
NP{que} VMOD{pode} VINF{abalar} NP{as crenças} AP{mais profundas} PP{de
quase} NP{toda a humanidade} ?}
```

Figure 20: Output of the infinitive adverbial rule (cataphora) (sentence (3.55))

```
MAIN(apontarem)                     MOD_POST(linha,pensamento)
VDOMAIN(apontarem,apontarem)        MOD_POST(pensamento,século XIX)
VDOMAIN(representam,representam)    MOD_POST(pensamento,naturalista)
MOD_PRE(descobertas,novas)          MOD_POST(século XIX,naturalista)
MOD_PRE(marco,novo)                 MOD_POST(teoria,evolução)
MOD_POST(pensamento,iniciada)       MOD_POST(apontarem,cura)
MOD_POST(descobertas,ciência)       MOD_POST(apontarem,doenças)
MOD_POST(marco,linha)               MOD_POST(representam,linha)
MOD_POST(marco,pensamento)          MOD_POST(representam,pensamento)
MOD_POST(marco,século XIX)          MOD_POST(representam,século XIX)
MOD_POST(autor,teoria)              MOD_POST(representam,naturalista)
MOD_POST(autor,evolução)            SUBJ_PRE(representam,descobertas)
MOD_POST(iniciada,século XIX)       CDIR_POST(representam,marco)
MOD_POST(iniciada,naturalista)      NE_INDIVIDUAL_PEOPLE(Charles Darwin)
MOD_POST(cura,doenças)              NE_T-ABSOLUT_TEMPO_DATE(em o século XIX)

3>TOP{A o VF{apontarem} PP{para a cura} PP{de doenças} , NP{as novas
descobertas} PP{de a ciência} VF{representam} NP{um novo marco} PP{em a
linha} PP{de pensamento} AP{iniciada} PP{em o NOUN{século XIX}} PP{por o
naturalista} NP{inglês} NP{NOUN{Charles Darwin}} , NP{autor} PP{de a
teoria} PP{de a evolução} .}
```

An interesting case happens when there is a chain of fronted subordinate clauses with zeroed NP subjects as in:

(3.56) *Ao* $Ø_i$ *processar estas frases sem ter antes criado as regras para* $Ø_i$ *desambiguar certas palavras, o sistema*$_i$ *irá certamente produzir erros*

'To process these sentences without before create the rules to disambiguate some words, the system will certainly produce errors'

In order to cope with (eventually) long strings of subclauses the general rule for fronted subordinate clauses has been expended by rule:

Figure 21: Rule for the infinitive adverbial subordinate clause (cataphora)

```
| ?*[verb], SC{?*, ?#1[verb,last]}, ?*[sc:~], SC{?*, ?#3[verb,last]} |

      if ( HEAD(#2,#1) & SUBORD(?,#2) & VDOMAIN(#2,#4) & SUBJ(#4,#5) &
HEAD(#6,#3) & & SUBORD(?,#6) VDOMAIN(#6,#7) & ~SUBJ(#7,?) )

      SUBJ[pre=+,anaph0=+](#7,#5)
```

In this rule two or more subordinate clauses without subjects can appear in the same string and the subject of the main verb is taken as the antecedent of the zeroed NP subjects.

3.5.5 Gerundive subordinate clause

Unlike infinitives (previous section), gerundive subordinate clauses do not have a conjunction to signal its subordinate status.

(3.57) *Essas mudanças*$_{i}$ *podem ser para o bem ou para o mal,* Ø$_{i}$ *atenuando sintomas de doenças ou* Ø$_{i}$ *provocando seu desenvolvimento*

'These changes can be for good or for evil, alleviating symptoms of disease or causing their development'

In fact, the gerund bound morpheme can be analyzed as the subordinate conjunction that links together the main and secondary clauses. Because of this, the semantic nexus between the two clauses is left undefined and directly depends on the meaning of each clause and our world knowledge.

Because of these differences, a specific rule was implemented for gerundive subordinate clauses which are very common in texts:

Figure 22: Rule for the gerundive subordinate clause

```
if ( MOD[post,gerund,sentential](#1,#7) & SUBJ[pre](#1,#5) & ~SUBJ(#7,?) )
    SUBJ[pre=+,anaph0=+](#7,#5)
```

The output of the sentence (3.57) is presented below:

Figure 23: Output of the gerundive subordinate rule (sentence (3.57))

```
MAIN(ser)                          MOD_SENTENTIAL_POST_GERUND(ser,atenuando)
COORD(ou,bem)                      MOD_SENTENTIAL_POST_GERUND
COORD(ou,mal)                      (atenuando,provocando)
VDOMAIN(atenuando,atenuando)       POSS_PRE(desenvolvimento,seu)
VDOMAIN(provocando,provocando)     SUBJ_PRE(ser,mudanças)
VDOMAIN(podem,ser)                 SUBJ_PRE_ANAPH0(atenuando,mudanças)
MOD_POST(sintomas,doenças)         SUBJ_PRE_ANAPH0(provocando,mudanças)
MOD_POST(ser,bem)                  CDIR_POST(atenuando,sintomas)
MOD_POST(ser,mal)                  CDIR_POST(provocando,desenvolvimento)

5>TOP{NP{Essas mudanças} VMOD{podem} VINF{ser} PP{para o bem} ou PP{para o
mal} , VGER{atenuando} NP{sintomas} PP{de doenças} ou VGER{provocando}
NP{seu desenvolvimento} .}
```

However, this rule heavily depends on previous parsing steps since gerundives often present subject inversion:

(3.58) *Esperando o governo*$_{i}$ *ganhar as eleições,* Ø$_{i}$ *lançou cá para fora novas leis eleitorais*

'Hoping the Government to win the election, [the Government] issued new electoral laws'

In this sentence, the subject of *esperando* 'expecting, hopping' is *o Governo* 'the Goverment'. Unless the correct subject dependency is extracted the anaphora will not be adequately resolved as it happened in this case.

3.5.6 Control verbs and nominal subordinate clauses

As it was mentioned in section 3.2.5, control verbs require a special set of rules to deal with the subcategorization constraints imposed by them, which have direct impact in zero anaphora resolution. One of the reasons for this is the fact that some of these nominal clauses can undergo syntactic restructuring and the subject of the dependent verb becomes, at surface, an autonomous constituent dependent of the main verb:

(3.59) *O Pedro mandou que a Ana lavasse a louça*

'Peter asked that Ana washed the dishes'

= *O Pedro mandou a Ana lavar a louça*

'Peter asked Ana to wash the dishes'

In this case, one does not want to consider that there is a zeroed NP subject anaphor of the infinitive since the subject of this verb is right next to it.

For a preliminary list of control verbs[44], a set of subcategorization features was defined:

- **s_inf**: the verb subcategorizes an infinitive and its subject is obligatorily coreferent to the zeroed subject in the infinitive;

(3.60) *O Pedro$_i$ prometeu Ø$_i$ lavar a louça*

'Peter promised to wash the dishes'

- **s_infdif**: the verb subcategorizes an infinitive and its subject cannot be coreferent to the zeroed subject in the infinitive;

(3.61) *O Pedro mandou lavar a louça*

'Peter ordered to wash the dishes'

- **s_np_inf**: the verb subcategorizes a direct object and an infinitive; the zeroed subject of the infinitive is obligatorily coreferent to the direct object;

(3.62) *O Pedro mandou a Ana$_i$ Ø$_i$ lavar a louça*

'Peter ordered Ana to wash the dishes'

- **s_np_ger**: the verb subcategorizes a direct object and a gerund; the zeroed subject of the gerund is obligatorily coreferent to the direct object;

(3.63) *O Pedro deixou a Ana$_i$ Ø$_i$ lavando a louça*

'Peter left Ana washing the dishes'

[44] This list has been initially compiled by Caroline Hagège and it integrates the Portuguese grammar (Mamede et al., 2010) developed under XIP. For the purpose of this dissertation, we expanded the feature set, added some few new verbs and revised the attributes for all the verbs of this list. At its current state, the list contains around 200 verbs (Appendix 5).

- **s_pp_inf**: the verb subcategorizes an indirect object and an infinitive; the zeroed subject of the infinitive is obligatorily coreferent to the indirect object;

(3.64) *O Pedro pediu à Ana*$_i$ *para Ø*$_i$ *lavar a louça*

'Peter asked to Ana for wash the dishes'

- **s_pp_qufconj**: the verb subcategorizes an indirect object and a finite subordinate clause in the subjunctive mode; the zeroed subject of the subordinate is obligatorily coreferent to the indirect object;

(3.65) *O Pedro pediu à Ana*$_i$ *que Ø*$_i$ *lavasse a louça*

'Peter asked to Ana that [she] wash the dishes'

- **s_qufconj**: the verb subcategorizes a finite subordinate clause in the subjunctive mode; the zeroed subject of the subordinate cannot be coreferent to the subject of the main clause;

(3.66) *O Pedro pediu que lavasse a louça*

'Peter asked that [someone] wash the dishes'

For example, the entries of verbs *prometer* 'promise' and *proibir* 'prohibit' in the XIP lexicon will look like this:

```
prometer:   verb += [s_inf:+].
proibir:    verb += [s_infdif:+,s_pp_inf:+, s_pp_qufconj:+, s_qufconj:+].
```

Since the general rules on infinitives would produce incorrect results in these cases, specific rules had already been developed to account for the subcategorization and coreferential constraints shown above. However the following rule has been added for verbs with `s_pp_qufconj` like *ordernar* 'to order':

(3.67) *O João ordenou à Ana*$_i$ *que Ø*$_i$ *lavasse a louça*

'John ordered to Ana that [she] washed the dishes'

In this case, the dative complement cannot be derived from the finite subordinate clause.

Figure 24: Rule for the control verbs

```
| #1[verb], ?*[verb:~], PP#8, SC{?*, ?#3[verb,last]} |
 if ( HEAD(#4[s_pp_qufconj],#1) & VDOMAIN(?,#4) & SUBJ(#4,#5) & HEAD(#6,#3)
& HEAD(#9,#8) & MOD[post](#4,#9) & VDOMAIN(#6,#7) & ~SUBJ(#7,?) )
     SUBJ[pre=+,anaph0=+](#7,#9)
```

The output of the sentence (3.67) is:

Figure 25: Output of the control verbs rule (sentence (3.67))

```
MAIN(ordenou)                    SUBJ_PRE_ANAPH0(lavasse,Ana)
VDOMAIN(ordenou,ordenou)         CDIR_POST(lavasse,louça)
VDOMAIN(lavasse,lavasse)         SUBORD(que,lavasse)
```

```
MOD_POST(ordenou,Ana)              NE_INDIVIDUAL_PEOPLE(Pedro)
SUBJ_PRE(ordenou,Pedro)            NE_INDIVIDUAL_PEOPLE(Ana)

8>TOP{NP{O Pedro} VF{ordenou} PP{a a Ana} SC{que VF{lavasse}} NP{a louça}
.}
```

3.5.7 Attributes

Adjectival constructions involve an auxiliary (copula) verb and give rise to a new binary dependency, ATTRIB[ute] between the subject and the adjective.

(3.68) *O Pedro estava alegre*

'Peter was happy'

In coordinate clauses, the subject of the second clause is reduced; therefore no subject dependency is extracted:

(3.69) *Ela um dia se casará e será muito infeliz*

'She will get married one day and will be very unhappy'

The copula verb also undergoes zeroing:

(3.70) *Branca de Neve*$_i$ *é* Ø$_i$ *tonta e boba por não haver se olhado no espelho — se olhou, não percebeu o fascínio e o terror que moram nele*

'Snow white is dumb and [is] silly for not having looked at herself on the mirror – she looked herself but did not notice the allure and the horror that live in it'

This happens because the subject dependency is formally defined as the element on which verbal agreement is expressed. Because of this, two rules were built. The first rule simply extends the anaphoric argument subject of the PREDSUBJ dependency to the ATTRIB dependency (Figure 26):

Figure 26: Rule for the attribute

```
if ( PREDSUBJ(#1[cop],#2) & SUBJ[anaph0](#1,#3) )
 ATTRIB[anaph0=+](#3,#2).
```

The second rule is slightly more complex for it checks on the other dependencies of the sentence without the copula verb in order to retrieve the subject anaphor dependecy (Figure 27):

Figure 27: Rule for the attribute

```
| #1[verb], ?*, CONJ[coord];PUNCT[lemma:";"];PUNCT[lemma:":"], (PP*;ADVP*),
AP#5 |

   if ( HEAD(#2,#1) & VDOMAIN(?,#2) & PREDSUBJ(#2,#3) & ATTRIB(#4,#3) &
HEAD(#6,#5)  & ~ATTRIB(?,#6) )

 ATTRIB[anaph0=+](#4,#6)
```

The ATTRIB_ANAPH0 feature extracted for sentence (3.69) were:

Figure 28: Output of the attribute rule (sentence (3.69))

```
MAIN(e)                          MOD_PRE(infeliz,muito)
COORD(e,casará)                  SUBJ_PRE(casará,Ela)
COORD(e,será)                    SUBJ_PRE_ANAPH0(será,Ela)
PREDSUBJ(será,infeliz)           ATTRIB(Ela,infeliz)
VDOMAIN(casará,casará)           ATTRIB_ANAPH0(Ela,infeliz)
VDOMAIN(será,será)               NE_TEMPO_DURATION(um dia)

115>TOP{NP{Ela} NP{NOUN{um dia}} NP{se} VF{casará} e VF{será} AP{muito
infeliz} .}
```

and for sentence (3.70) were:

Figure 29: Output of the attribute rule (sentence (3.70))

```
MAIN(e)                          MOD_PRE_NEG(olhado,não)
COORD(e,fascínio)                SUBJ_PRE(é,Branca de Neve)
COORD(e,tonta)                   SUBJ_PRE_ANAPH0(olhado,Branca de Neve)
COORD(e,terror)                  SUBJ_PRE_ANAPH0(percebeu,Branca de Neve)
COORD(e,boba)                    SUBJ_PRE_ANAPH0(olhou,Branca de Neve)
PREDSUBJ(é,tonta)                SUBJ_PRE_ANAPH0(moram,Branca de Neve)
VDOMAIN(é,é)                     CDIR_POST(percebeu,fascínio)
VDOMAIN(olhou,olhou)             CDIR_POST(percebeu,terror)
VDOMAIN(percebeu,percebeu)       SUBORD_CAUSA(por,haver)
VDOMAIN(moram,moram)             SUBORD(que,moram)
VDOMAIN(haver,olhado)            SUBORD(se,olhou)
MOD_POST(olhado,espelho)         ATTRIB(Branca de Neve,tonta)
MOD_POST(percebeu,ele)           ATTRIB_ANAPH0(Branca de Neve,boba)
MOD_POST(moram,ele)              NE_INDIVIDUAL_PEOPLE(Branca de Neve)
MOD_PRE_NEG(percebeu,não)

118>TOP{NP{NOUN{Branca de Neve}} VCOP{é} AP{tonta} e AP{boba} SC{por
ADVP{não} VTEMP{haver}} NP{se} VPP{olhado} PP{em o espelho} - SC{se
VF{olhou}} , ADVP{não} VF{percebeu} NP{o fascínio} e NP{o terror} SC{que
VF{moram}} PP{em ele} .}
```

4 Evaluation: Results and discussion

This chapter presents results from the application of the XIP parser, enriched with new rules described above and a brief discussion on the main errors is also made.

4.1 Results

In order to evaluate the performance of the parser with the rules described above, the evaluation corpus was split in sentences and only sentences that present zero anaphors cases were selected[45]. The evaluation corpus contained 235 zero anaphors in 174 sentences. Then the output of the parser was manually verified.

Results are expressed using the measures of Precision (P), Recall (R) and F-measure[46] and they are presented in Table 8.

Table 8: Zero anaphora rules results

Measures	Results	%
Precision	0.6011	60.11%
Recall	0.4553	45.53%
F-measure	0.5181	51.81%

These results, while not yet satisfactory, are encouraging, specifically when one takes into consideration that this is likely the first attempt at a rule-based ZAR in (Brazilian) Portuguese.

In the next section (4.2), the most common errors will be presented and discussed. Our general goal is to identify the problems found in the ZAR task. These problems fall mainly on three types of errors sometimes connected: (a) POS tagging, (b) chunking and (c) dependency extraction, including ZA rules.

[45] The impersonal, indefinites, indefinite first person plural, third person plural and cataphoras were not considered.

[46] The precision measure is calculated considering the total number of correct cases (i.e. the cases in which the parser correctly assigned the `ANAPH0` feature) divided by the total number of the `ANAPH0` feature assigned by the parser (which includes the cases in which the feature was mistakenly assigned). The recall measure is calculated considering the total number of correctly cases identified divided by the total number of zero anaphora annotated on the corpus. F-measure is the harmonic mean of P and R: 2·P·R/P+R.

4.2 Discussion

4.2.1 Errors from POS tagger

Consider the following sentence:

(4.1) *A Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior (Capes)$_i$ fará uma avaliação especial dos mestrados profissionais, Ø$_i$ levando em conta suas especificidades*

'The Coordenação de Aperfeiçoamente de Pessoal do Ensino Superior (Capes) will perform a special evaluation of the profissional masters taking into consideration their specificities'

where the following output was extracted:

Figure 30: POS tagger errors (sentence (4.1))

```
MAIN(Capes)                          MOD_POST(fará,mestrados)
VDOMAIN(Capes,Capes)                 MOD_SENTENTIAL_POST_GERUND(fará,levando)
VDOMAIN(fará,fará)                   POSS_PRE(especificidades,suas)
VDOMAIN(levando,levando)             SUBJ_PRE(Capes,Coordenação
MOD_POST(avaliação,especial)         de Aperfeiçoamento
MOD_POST(mestrados,profissionais)    de Pessoal de o Ensino Superior)
MOD_POST(avaliação,mestrados)        CDIR_POST(fará,avaliação)
MOD_POST(especial,mestrados)         CDIR_POST(levando,especificidades)

2>TOP{NP{A NOUN{NOUN{Coordenação de Aperfeiçoamento} de NOUN{Pessoal de o
Ensino} Superior}} ( VF{Capes} ) VF{fará} NP{uma avaliação} AP{especial}
PP{de os mestrados} AP{profissionais} , VGER{levando} ADVP{em conta}
NP{suas especificidades} .}
```

The parser considered the acronym *Capes* as a verb (verb *capar* 'to castrate', *capes*_2ndsg). As a result, the subject of the verb *fará* 'will perform' was not properly identified and therefore the subject of the verb in the gerundive (*levando* 'taking') was also misidentified.

4.2.2 Errors due to the shallow parser

Consider the follow sentence:

(4.2) *Zé Galego$_i$, de 48 anos, cujo pai veio do Ceará para o Acre, está começando um negócio de comércio, Ø$_i$ levando os produtos da floresta para a sede do município e trazendo mercadorias*

'Zé Galego, 48 years, whose father came from Ceará to Acre, is starting a business, [he] is taking products from the forest to the county headquarter and [he] is bringing products'

where the following output was extracted:

Figure 31: Shallow parser errors (sentence (4.2))

```
MAIN(veio)                    MOD_POST(veio,Acre)
VDOMAIN(veio,veio)            MOD_POST(levando,sede)
VDOMAIN(levando,levando)      MOD_POST(levando,município)
VDOMAIN(trazendo,trazendo)    MOD_SENTENTIAL_POST_GERUND(começando,levando)
VDOMAIN(está,começando)       MOD_SENTENTIAL_POST_GERUND(levando,trazendo)
MOD_POST(negócio,comércio)    SUBJ_PRE(veio,pai)
MOD_POST(produtos,floresta)   SUBJ_PRE_ANAPH0(trazendo,pai)
MOD_POST(produtos,sede)       CDIR_POST(começando,negócio)
MOD_POST(produtos,município)  CDIR_POST(levando,produtos)
MOD_POST(Ceará,Acre)          CDIR_POST(trazendo,mercadorias)
MOD_POST(floresta,sede)       NE_INDIVIDUAL_PEOPLE(Zé Galego)
MOD_POST(floresta,município)  NE_LOCAL_ADMIN_AREA(Ceará)
MOD_POST(sede,município)      NE_LOCAL_ADMIN_AREA(Acre)
MOD_POST(veio,Ceará)          NE_QUANT(48 anos)

55>TOP{" NP{NOUN{Zé Galego}} , PP{de NOUN{48 anos}} , cujo NP{pai} VF{veio}
PP{de o Ceará} PP{para o Acre} , VASP{está} VGER{começando} NP{um negócio}
PP{de comércio} , VGER{levando} NP{os produtos} PP{de a floresta} PP{para a
sede} PP{de o município} e VGER{trazendo} NP{mercadorias} .}
```

Probably because of the insertion and the relative clause, the parser failed to extract the SUBJ dependency between *Zé Galego* and *está começando* ‘is starting’, and therefore the remaining anaphoric subject of the gerundive *levando* ‘is taking’ and *trazendo* ‘is bringing’ were not adequately identified.

4.2.3 Errors due to inadequate processing of the relative clauses

Consider the following sentence:

(4.3) Luiz também foi atacado, quase simultaneamente, pelo mesmo ser, mas foi salvo pelo telescópio, que foi arrebatado de sua mão e ganhou as alturas, caindo já aos pedaços metros à frente

‘Luis was also attacked, almost simultaneously, for the same creature, but [he] was saved by the telescope, which was taken from his hand and it was thrown up, beat-up yards ahead’

where the following output was extracted:

Figure 32: Relative clause errors (sentence(4.3))

```
MAIN(atacado)                    MOD_SENTENTIAL_POST_GERUND(ganhou,caindo)
VDOMAIN(ser,ser)                 POSS_PRE(mão,sua)
VDOMAIN(ganhou,ganhou)           SUBJ_PRE(atacado,Luiz)
VDOMAIN(caindo,caindo)           SUBJ_PRE(ser,mesmo)
VDOMAIN(foi,atacado)             SUBJ_PRE_ANAPH0(salvo,Luiz)
VDOMAIN(foi,salvo)               SUBJ_PRE_ANAPH0(ganhou,Luiz)
VDOMAIN(foi,arrebatado)          SUBJ_PRE_ANAPH0(arrebatado,Luiz)
MOD_PRE(atacado,também)          SUBJ_PRE_ANAPH0(caindo,Luiz)
MOD_PRE(ser,simultaneamente)     CDIR_POST(ganhou,alturas)
```

```
MOD_PRE(simultaneamente,quase)     CDIR_SENTENTIAL_POST_INF(atacado,ser)
MOD_POST(salvo,telescópio)         SUBORD(que,foi)
MOD_POST(arrebatado,mão)           NE_INDIVIDUAL_PEOPLE(Luiz)
MOD_POST(caindo,pedaços)           NE_QUANTITY_QUANT(metros)
MOD_POST(caindo,frente)

165>TOP{NP{Luiz} ADVP{também} VCOP{foi} VCPART{atacado} , ADVP{quase
simultaneamente} , PP{por o mesmo} VINF{ser} , mas VCOP{foi} VCPART{salvo}
PP{por o telescópio} , SC{que VCOP{foi}} VCPART{arrebatado} PP{de sua mão}
e VF{ganhou} NP{as alturas} , VGER{caindo} ADVP{já} PP{a os pedaços}
NP{metros} PP{a a frente} .}
```

In this sentence, the relative clause has not been correctly identified, since *que* 'which' was considered a conjunction. Because of that, the parser attributed a SUBJ_PRE_ANAPH0 to the verb of the relative clause *que foi arrebatado* 'which was taken'. The remaining error results from this. Also, the POS tagger failed to tag *ser*[47] as verb.

4.2.4 Errors due to lack of information in the lexicon

Consider the following sentence:

(4.4) *O Pedro*$_i$ *pediu à Ana que Ø*$_i$ *lavasse a louça*

'Peter asked to Ana that [she] washed the dishes'

where the following output were extracted:

Figure 33: Lack of information in the lexicon (sentence (4.4))

```
MAIN(pediu)                        SUBJ_PRE_ANAPH0(lavasse,Pedro)
VDOMAIN(pediu,pediu)               CDIR_POST(lavasse,louça)
VDOMAIN(lavasse,lavasse)           SUBORD(que,lavasse)
MOD_POST(pediu,Ana)                NE_INDIVIDUAL_PEOPLE(Pedro)
SUBJ_PRE(pediu,Pedro)              NE_INDIVIDUAL_PEOPLE(Ana)

8>TOP{NP{O Pedro} VF{pediu} PP{a a Ana} SC{que VF{lavasse}} NP{a louça} .}
```

The parser incorrectly assign the SUBJ_PRE_ANAPH0(lavasse,Pedro) dependency in which the NP *Pedro* 'Peter' is assigned to the verb *lavasse* 'washed' because on the list of control verbs is missing the information that the verb *pedir* 'to ask' is s_pp_qufconj what means that the zeroed subject of the subordinate is obligatorily coreferent to the indirect object.

[47] In Portuguese the word *ser* is ambiguous. It can be either a *creature* or the infinitve form of the verb *to be*.

4.2.5 Errors due to ambiguity between adjectives and past participles

Consider the following sentence:

(4.5) *O Pedro*$_i$ *estava cansado mas* Ø$_i$ *não estava exausto*

'Peter was tired but [he] was not exhausted'

where the following output were extracted:

Figure 34: Adjectives/Past Particples error analizes (sentence (4.5))

```
MAIN(cansado)                    MOD_PRE_NEG(estava,não)
VDOMAIN(estava,estava)           SUBJ_PRE(cansado,Pedro)
VDOMAIN(exausto,exausto)         SUBJ_PRE_ANAPH0(estava,Pedro)
VDOMAIN(estava,cansado)          NE_INDIVIDUAL_PEOPLE(Pedro)

114>TOP{NP{O Pedro} VCOP{estava} VCPART{cansado} mas ADVP{não} VF{estava}
VF{exausto} .}
```

The parser should to assign the ATTRIB_ANAPH0 dependency to the word exausto 'exhausted' but it was not assign because this word is ambiguous (past participle or adjective).

4.2.6 NP assigned incorrectly

Consider the following sentence:

(4.6) *Uma lufada de ar frio entrou pela janela*$_i$ *quando* Ø$_i$ *foi aberta*

'A blast of cold air came through the window when [it] was opened'

where the following output were extracted:

Figure 35: Incorrect NP assigned (sentence (4.6))

```
MAIN(entrou)                     MOD_POST(lufada,ar)
VDOMAIN(entrou,entrou)           MOD_POST(entrou,ela)
VDOMAIN(foi,aberta)              SUBJ_PRE(entrou,lufada)
MOD_POST(ar,frio)                SUBJ_PRE_ANAPH0(aberta,lufada)
SUBORD(quando,foi)

158>TOP{NP{Uma lufada} PP{de ar} AP{frio} VF{entrou} PP{por ela} SC{quando
VCOP{foi}} VCPART{aberta} .}
```

In this case, the subject of the subordinate clause introduced by *quando* 'when' is in a previous PP. As there is no explicit subject, the general rule was applied[48]. In order to bypass these general rules, information on distributional constraints would be needed.

As general remark, when subordinate clauses do not take the subject of the main clause as their subjects' antecedent, the ZAR rules fail. Other strategies must, therefore, be found to solve these cases.

[48] In EP this zeroing is hardly acceptable.

5 Conclusion and future work

The objectives of this dissertation were achieved: we presented a systematic linguistic analysis of syntactic constraints on zero anaphora in (Brazilian) Portuguese, a typical syntactical structure of this language, and produced a set of linguistically motivated rules to endow the rule-base parser XIP to resolve zero subject anaphora in a fully integrated NLP chain (Mamede et al., 2010).

To this end, a specific corpus, the ZAC corpus (Pereira, 2009) has been built, including different textual genres. All texts that compose the corpus were taken from the Brazilian Portuguese variety. A set of sentences, some retrieved from the NILC corpus and other especially constructed to test zero anaphora resolution (ZAR) rules, was also put together. These sentences were collected/formed in order to have examples of a varied set of situations in which the zero anaphora phenomena occurs and to work as a testbed for the ZAR rules.

The corpus was divided in two parts, one for the training and the other for the testing phase. The corpus and the sentences were manually annotated. A set of annotation guidelines was provided to ensure good annotation reproducibility. The test corpus was independently annotated by a linguist using the same guidelines that were previously discussed and defined.

Rules were developed based on the analysis of syntactical and semantic structures of sentences selected and also using our intuition as native speakers of the language. The zero anaphora cases were limited to investigate zeroed NP subject within the same sentence (intrasentential anaphora). Although some cataphora rules were implemented, these cases were not considered when analyzing and calculating the final results of the implementation rules.

Rules were implemented in order to enable the XIP parser to recover zeroed NP subjects based on a previously defined grammar implemented in this parsing system. In particular, the ZAR rules rely on the previous processing steps of the NLP chain (Mamede et al., 2007), namely, a tokenizer, a POS tagger, a rule-base POS disambiguation module, and the XIP parser proper, which performs the chunking of the sentences and extract syntactic-semantic dependencies among chunks. Results on the ZAR rules' module, which are the last step of the parser's processing, are, therefore, dependent on the results of these previous modules of the NLP chain.

Results are promising: the system attains a 60.11% Precision, 45.53% Recall

and a F-measure of 51.81%. In spite of these results, much is still left to be done, foremost the improvement of Precision. In the discussion of this results, it was possible to verify that some errors came from incorrect POS tagging and/disambiguation.

The most important errors, however, result from insufficient development of the dependencies rules: they still are not performed enough to capture all explicit subjects, particularly in subordinate (adverbial and nominal) and relative clauses, thus precluding the recovery of zeroed anaphors.

Finally, even if the ZAR rules were built having in mind the Brazilian variety of Portuguese, it became evident from our experiments that the European variety only seldom differs from the American, hence much work is expected be reusable.

5.1 Future work

As it was mentioned above, the performance of the parser can be improved with the adequate processing of some syntactic structures. The correct identification of the relative clauses is one of the cases that require significant improvement; if the parser correctly identifies the syntactic function of the relative pronouns then errors due to incorrectly assignment of a zeroed subject in the relatives (and other subsequent clauses) will not occur.

Information on distributional constraints can also improve the performance of the rules. If distributional constraints on verbal arguments are provided in the lexicon, then the selection of the antecenden NP will be improved, especially in the cases where the zeroed NP subjects is in a PP chunk or integrated in a subordinate clause instead of the subject position of main verbs.

The correct identification of impersonal verbs, indefinite first and third person-plural and the indefinite -se pronoun constructions are also a research area of some importance for ZAR, because the zero anaphor would not be wrongly assigned in this cases.

Finally, the treatment of the cataphora could also improve the zero anaphora resolution. If a sentence is composed by a sequence of clauses in which the first clause presents a zeroed cataphoric subject, then the correctly identification of the zeroed NP subject of the first clause can help the recovery of the subject of subsequent clauses.

In view of these tasks, the first step will be to prepare the system to improve the resolution of intrasentential zeroed subjects, especially by refining the relative clause dependencies' extraction. Next, using information already available on verbal, nominal and adjectival distributional constraints, we intent to refine the ZAR module, by recovering zero anaphors whose antecendents are in the previous discourse. For this, average distance between intersentential anaphor and their antecedents can already be computed from the ZAC corpus to guide the heuristics.

A systematic comparison between the two main varieties of Portuguese is also in order, in particular in order to establish the particular differences that distinguish them. It is expected that results on this area will have significant impact on several applications (for example in text generation).

Naturally, much is still left to be done, but we expect to have contributed to a better understanding of the complexity of zero anaphora in Portuguese.

References

Ait-Mokhtar, S.; Chanod, J.; Roux, C. 2002. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering 8* (2/3). London, Cambridge University Press. pp 121-144. Available at: http://journals.cambridge.org/action/displayFulltext?type=1&fid=116936&jid=NLE&volumeId=8&issueId=2-3&aid=116935

Aone, C.; McKee, D. 1993. A language-independent anaphora resolution system for under-standing multilingual texts. *Proceedings of the 31st Annual Meeting of the ACL* (ACL. 95). Colimbus, OH. pp: 156-163. Available at: http://www.aclweb.org/anthology/P/P93/P93-1021.pdf

Aone, C.; Bennett, S. 1994. Discourse tagging tool and discourse-tagged multilingual corpora. *Proceeding of the International Workshop on Sharable Natural Language Resources* (SNLR). Nara, Japan. pp: 71-77. Available at: http://academic.research.microsoft.com/Paper/295562.aspx

Aone, C.; Bennett, S. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. *Proceedings of the 33rd Annual Meeting of the ACL*. (ACL. 95). Cambridge, Mass. pp: 122.129. Available at: http://www.aclweb.org/anthology-new/P/P95/P95-1017.pdf

Aone, C.; Bennett, S. 1996. Applying machine learning to anaphora resolution. In Wermter, S.; Riloff, E.; Scheler,G. (Eds.) *Connectionist, statistical and symbolic approaches to learning for Natural Language Processing*. Berlin: Springer. pp: 302-314. Available at: http://www.springerlink.com/content/fq15236388hhk423/fulltext.pdf

Baldwin, B. 1997. CogNIAC: high precision coreference with limited knowledge and linguistics resources. *Proceedings of the ACL97/EACL97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*. Madrid. pp: 38-45. Available at: http://www.aclweb.org/anthology/W/W97/W97-1306.pdf

Baptista, J.; Mamede, N.; Gomes, F. 2010. Auxiliary Verbs and Verbal Chains in European. In Pardo, T. *et al.* (Eds.): *PROPOR 2010, LNAI 6001*. pp: 110-119.

Bechara, I. 2001. *Moderna gramática portuguesa*. Rio de Janeiro: Lucerna.

Bick, E. 2000. *The parsing system PALAVRAS: Automatic grammatical analysis of Portuguese in a constraint grammar framework*. PhD thesis, Arthus University, Denmark. Available at: http://beta.visl.sdu.dk/~eckhard/pdf/PLP20-amilo.ps.pdf

Brito, A.; Duarte, I.; Matos, G. 2003. Estrutura da frase simples e tipos de frase. In: *Gramática da Língua Portuguesa.* Lisboa: Caminho. pp 442-449.

Carbonell, J.G.; Brown, R. 1988. Anaphora resolution: a multi-strategy approach. *Proceedings of the 12th International Conference on Computational Linguistics* (COLING'88). Budapest: Hungary. pp: 96-101. Available at: http://www.cs.cmu.edu/~jgc/publication/Anaphora_Resolution_A_Multi_Strategy_ICCL_1988.pdf

Cardie, C. 1992. Learning to disambiguate relative pronouns. *Proceedings of the 10th National Conference on Artificial Intelligence* (AAAI'92). San Jose, Calif. pp: 38-43. Available at: http://www.cs.cornell.edu/home/cardie/papers/aaai-92.pdf

Cardie, C.; Pierce, D. 1998. Error-driven pruning of treebank grammars for base noun phrase identification. *Proceedings of the 36th Annual Meeting of the ACL and COLING-98*. Montreal, Canada. pp: 218-224. Available at: http://www.aclweb.org/anthology/P/P98/P98-1034.pdf

Cardie, C.; Wagstaff, K. 1999. Noun phrase coreference as clustering. *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*. University of Maryland, USA. pp: 82-89. Available at: http://acl.ldc.upenn.edu/W/W99/W99-0611.pdf

Carter, D. 1986. *A shallow processing approach to anaphor resolution*. PHD thesis, University of Cambridge.

Carvalho, A.; maduro, R. 2002. Syntactic Analysis for Ellipsis Handling in Coordinated Clauses. In: *16th Brazilian Symposium on Artificial Inteligence (SBIA 2002)*. Porto de Galinhas, Recife. pp: 397-406. Available at: http://www.lsi.us.es/iberamia2002/confman/SUBMISSIONS/180-ritorizmar.pdf

Chaves, A.; Rino, L. 2008. "The Mitkov Algorithm for Anaphora Resolution in Portuguese". *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings* (PROPOR 2008). Aveiro, Portugal. pp. 51-60.

Chomsky, N. 1957. *Syntactic Structures*. Paris; The Nerdlands: Mouton & CO.

Chomsky, N. 1981. *Lectures on government and binding*. Berlin; New York: Mouton de Gryter.

Coelho, T. 2005. *Resolução de anaphora pronominal em português utilizando o algorítmo de Lappin and Leass*. Master dissertation, Universidade Estadual de Campinas. Campinas, São Paulo. Available at: http://libdigi.unicamp.br/document/?code=vtls000390497

Collovini, S.; Carbonel, T.; Fuchs, J.; Coelho, J.; Rino, L.; Vieira, R. 2007. Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. *Anais do XXVII Congresso da SBC TIL V Workshop em Tecnologia da Informação e da Linguagem Humana.* Rio de Janeiro. pp: 1605-1614

Cuevas, R.; Honda, W.; Lucen, D.; Paraboni, I.; Oliveira, P. 2008. Portuguese Pronoun Resolution: Resources and Evaluation. *9th International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing-2008). Haifa, Israel. pp: 344-350. Available at: http://each.uspnet.usp.br/ivandre/papers/cicling2008.pdf

Cunha, C.; Cintra, L. 1984. *Nova Gramática do Português Contemporâneo*. Lisboa: Edições João Sá da Costa.

Dagan, I.; Itai, A. 1990. Automatic processing of large corpora for the resolution of anaphora reference. *Proceedings of the International Conference on Computational Linguistics* (COLING 90). Helsink. pp: 1-3. Available at: http://www.aclweb.org/anthology/C/C90/C90-3063.pdf

Duarte, I. 2003. Subordinação completive – as orações completivas. In: Mateus *et al*. 2003. pp: 595-651.

Evans, R. 2000. A comparison of rule-based and machine learning methods for identifying non-nominal it. In *Proceedings of NLP 2000*. Patras, Greece. pp: 233-241.

Ferrandez, A.; Peral, J. 2000. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-00)*. Hong Kong. pp: 166-172. Available at: http://www.aclweb.org/anthology-new/P/P00/P00-1022.pdf

Ferrández, A.; Palomar, M.; Moreno, L. 1998. Anaphora resolution in unrestricted texts with partial parsing. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (COLING-ACL'98). Montreal, Canada. pp: 385-391. Available at: http://portal.acm.org/citation.cfm?id=980911&coll=GUIDE&dl=GUIDE&CFID=79491104&CFTOKEN=22371270&ret=1#Fulltext

Ferrández, A.; Palomar, M.; Moreno, L. 1999. An empirical approach to Spanish anaphora resolution. *Machine Translation*, 14(3/4). pp: 191-216. Available at: ftp://altea.dlsi.ua.es/people/antonio/ART_MUL5.pdf

Gasperin, C.; Vieira, R.; Goulart, R.; Quaresma, P. 2003. Extracting XML syntactic chunks from Portuguese corpora. *Proceedings of the Workshop on Traitement automatique des langues minoritaires*. Bartz-sur-Mer. pp: 223-232. Available at: http://www.rodrigo.goulart.nom.br/publicacoes/gasperin2003a.pdf

Ge, N.; Hale, J.; Charniak, E. 1998. A statistical approach to anaphora resolution. *Proceedings of the Workshop on Very Large Corpora*. Montreal, Canada. pp: 161-170. Available at: http://www.aclweb.org/anthology/W/W98/W98-1119.pdf

Gross, M. 1975. On the relations between syntax and semantics. In E. L. Keenan (Ed.), *Formal semantics of natural language*. Cambridge: Cambridge University Press. pp: 389–405

Hagège, C.; Baptista, J.; Mamede, N. 2008. Reconhecimento de entidades mencionadas com o XIP: Uma colaboração entre a Xerox e o L2F do INESC-ID Lisboa. In Mota, C.; Santos, D. (Orgs) *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Aveiro: Linguateca. pp: 261-274. Available at: http://www.inesc-id.pt/pt/indicadores/Ficheiros/5759.pdf

Hagège, C.; Baptista, J.; Mamede, N. 2010. Caracterização e Processamento de Expressões Temporais em Português. Linguamática 2-1. pp: 63-77.

Halliday, M.; Hasan, R. 1976. *Cohesion in English*. London: Longman.

Harabagiu, L.; Maiorano, S. 1999. Knowledge-learn conference resolution and its relation to textual cohesion and conference. *Proceeding of the ACL 99 Workshop on the Relation of Discourse/Dialogue Structure and Reference*. College Park, Md. pp: 29-38. Available at: http://www.aclweb.org/anthology/W/W99/W99-0104.pdf

Harris, Z. 1981. *Papers on Syntax*. Henry Hiz (Ed.). Dordrecht: D.Reidel Publishing Company.

Harris, Z. 1991. *A Theory of Language and Information: A mathematical approach*. Oxford: Clarendon Press.

Hobbs, J. 1976. *Pronoun resolution*. Research Report 76-1. New York: Department of Computer Science, City University of New York.

Hobbs, J. 1978. Resolving pronoun references. *Lingua*, 44. pp: 339-352. Available at: http://www.isi.edu/~hobbs/ResolvingPronounReferences.pdf

Huddleston, R.; Pullum, G. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Kennedy, C.; Bougarev, B. 1996. Anaphora for everyone: pronominal anaphora resolution without a parser. *Proceedings of the International Conference on Computational Linguistics* (COLING 96). Copenhagen. pp: 113-118. Available at: http://www.aclweb.org/anthology/C/C96/C96-1021.pdf

Lappin, S.; Leass, H. 1994. An Algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4). pp: 535-561. Available at: http://www.aclweb.org/anthology/J/J94/J94-4002.pdf

Lee T.; Lewicki, M.; Sejnowski, T. 2000. ICA mixture models for unsupervised classification of non-Gaussian classes and automatic context switching in blind signal separation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 22, nº 10. pp: 1078–1089. Available at: http://www.cnbc.cmu.edu/cplab/papers/lee-lewicki-sejnowski-00.pdf

Mamede, N.; Baptista, J.; Vaz, P.; Hagège, C. 2010. *Nomenclature of chunks and dependencies in Portuguese XIP grammar* (v. 2.1.). Internal Report. Lisboa: L2F/INESD-ID Lisboa.

Marcus, M.; Santorini, B.; Marcinkiewicz, M.A. 1993. Building a large annotated corpus of Englih: the Penn Treebank. *Computational Linguistics*, 19 (2). pp: 313-330. Available at: http://acl.ldc.upenn.edu/J/J93/J93-2004.pdf

Mateus, M.; Brito, A.; Duarte, I.; Faria, I.; Frota, S.; Matos, G.; Oliveira, F.; Vigário, M.; Villalva, A. 2003. *Gramática da Língua Portuguesa*. Lisboa: Caminho.

Matos, G. 2003a. Estruturas de coordenação. In: Mateus *et al*. 2003: pp: 551-592.

Matos, G. 2003b. Construções elípticas. In: Mateus *et al.* 2003: pp: 869-913.

McCarthy, J.; Lehnert, W. 1995. Using decision trees for coreference resolution. *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (IJCAI). Montreal, Canada. pp: 1050-1055. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.1737

Mendes, A.; Coheur, L.; Mamede, N.; Ribeiro, R.; Batista, F.; Matos, D. 2007. QA@L2F, first steps at QA@CLEF. In CLEF 2007 Proceedings, Lecture Notes in Computer Science. Springer, 2008. pp: 356-363. Available at: http://www.inesc-id.pt/pt/indicadores/Ficheiros/4918.pdf

Mitkov, R. 1998a. Evaluating anaphora resolution approaches. *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium* (DAARC 2). Lancaster. pp: 164-172.

Mitkov, R. 1998b. Robust pronoun resolution with limited knowledge. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*. Montreal, Canada. pp: 869-875.

Mitkov, R. 2000. Pronoun resolution: the practical alternative. In Botley, S.; McEnery, A. (Eds), *Discourse Anaphora and Anaphor Resolution*. Amsterdam: John Benjamin Publishers. pp: 129-143.

Mitkov, R. 2002. *Anaphora Resolution*. London: Longman.

Mitkov, R. 2003. Anaphora Resolution. In: Mitkov, R (Ed), 2003. *The Oxford Handbook of Computational Linguistics.* New York: Oxford University Press. pp 266-283.

Mitkov, R. Barbu, C. 2000. Improving pronoun resolution in two languages by means of bilingual corpora. *Proceedings of the Discourse, Anaphora and Reference Resolution Conference* (DAARC2000). Lancaster. pp: 133-137.

Mitkov, R.; Belguith, L.; Stys, M. 1998. Multilingual robust anaphora resolution. *Proceedings of the 3rd International Conference on Empirical Methods in Natural Language Processing* (EMNLP-3). Granada. pp: 7-16.

Mitkov, R.; Orasan, C.; Evans, R. 1999. The importance of annotated corpora for NLP: the cases of anaphora resolution and clause splitting. In: *Proceedings of TALN'99*. Corsica, France. pp. 60-69. Available at: http://clg.wlv.ac.uk/papers/mitkov-99b.pdf

Mitkov, R., Evans, R. and Orasan, C. 2002. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Proceedings of CICLing-2000*. Mexico City, Mexico. pp: 168-186. Available at: http://clg.wlv.ac.uk/papers/ciclingAR19.pdf

Molinier, C.; Lévrier, F. 2000. Grammaire des Adverbes: Description des Formes en –ment. Genève: Librairie Droz.

Mooney, R. 2003. Machine Learning. In: Mitkov, R (Ed), 2003. *The Oxford Handbook of Computational Linguistics.* New York: Oxford University Press. pp 376-394.

Mota, C.; Santos, D. (Eds.). 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca. Available at: http://www.linguateca.pt/LivroSegundoHAREM/

Nakaiwa, H. 1997. Automatic Identification of Zero Pronouns and their Antecedents within Aligned Sentence Pairs. pp: 127-141. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.6748

Nakaiwa, H.; Ikehara, S. 1995. Intrasentential Resolution of Japanese Zero Pronouns in a Machine Translation System using Semantic and Pragmatic Constraints. pp: 96-105. Available at: http://www.mt-archive.info/TMI-1995-Nakaiwa.pdf

Nasakawa, T. 1994. Robust method of pronoun resolution using full-text information. *Proceedings of the 15th International Conference on Computational Linguisticas* (COLING 94). Kyoto. pp: 1157-1163.

Ng, V.; Cardie, C. 2002. Improving machine learning approaches to coreference resolution. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (Acl' 02). Philadelphia, Pa. pp: 104-111.

Oliveira, P.; Romero, R. 2004 Enhanced ICA Mixture Model for Unsupervised Classification. *9th Ibero-American Conference on Artificial Intelligence* IBERAMIA 2004. pp: 205-214.

Orasan, C. 2000. CLinkA a Coreferential Links Annotator. In *Proceedings of LREC'2000*. Athens, Greece. pp: 491–496. Available at: http://clg.wlv.ac.uk/papers/orasan-00b.pdf

Orasan, C.; Evans, R.; Mitkov, R. 2000. Enhancing preference-based anaphora resolution with genetic algorithms. *Proceedings of NLP 2000*. Patras, Greece. pp: 185-195. Available at: http://clg.wlv.ac.uk/papers/orasan-NLP-00.pdf

Palomar, M.; Ferrández, A.; Moreno, L.; Saiz-Noeda, M.; Muñoz, R.; Martinez-Barco, P.; Peral, J.; Navarro, B. 1999. A robust partial parsing strategy based on the slot unification grammars. *Proceedings of the 6th Conference on Natural Language Processing* (TALN'99). Corsica, France. pp: 263-272. Available at: http://www.atala.org/doc/actes_taln/AC_0129.pdf

Palomar, M.; Ferandez, L.; Martínez-Barco, P.; Peral, J.; Saiz-Noeda, M.; Muñoz, R. 2001. An algorithm for anaphora resolution in Spanish texts. *Computational Linguistics*, 27(4). pp: 545-567. Available at: http://www.aclweb.org/anthology-new/J/J01/J01-4005.pdf

Pereira, S. 2009. ZAC.PB: An Annotated Corpus for Zero Anaphora Resolution in Portuguese. In Student Research Workshop Proceedings held in conjunction with The International Conference RANLP. Borovets, Bulgaria. pp: 53-59. Available at: http://lml.bas.bg/ranlp2009/DOCS/ranlp2009_W7.pdf#page=61

Peters, C.; Jijkoun, V.; Mandl, Th.; Müller, H.; Oard, D.W.; Peñas, A.; Petras, V.; Santos, D. (Eds.). 2007. 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007. Budapest, Hungary. Series: Lecture Notes in Computer Science, Vol. 5152.

Pinheiro, G.; Aluísio, S. 2003. *Corpus NILC: descrição e análise crítica com vistas ao projeto Lacio-Web.* Série de Relatórios Técnicos do Instituto de Ciencias Matemáticas e de Computação – ICMC, Universidade de São Paulo, N. 190. Available at: http://www.linguateca.pt/CETENFolha/

Quinlan, J. 1993. *C4.5: Programs for machine learning*. San Mateo, California: Morgan Kaufmann.

Quirk, R.; Greenbaum, S.; Leech, G.; Svartvik, J. 1985. *A Comprehensive Grammar of the English Language (General Grammar)*. London: Longman.

Ranchhod, E. 1990. *Sintaxe dos Predicados Nominais com* ***estar***. Lisboa: INIC. pp 77.

Reinhart, T. 1983. *Anaphora and semantic interpretation*. London: Croom Helm.

Rello, L.; Ilisei, I. 2009. A Comparative Study of Spanish Zero Pronoun Distribuition. *Besançon: International Symposium on Data and Sense Mining, Machine Tanslation and Controlled Languages*. Besançon, French. pp. 209-214. Available at: http://clg.wlv.ac.uk/papers/Ilisei_ZP-ISMTCL.pdf

Rich, E; LuperFoy, S. 1988. An architecture for anaphora resolution. *Proceedings of the Second Conference on Applied Natural Language Processing* (ANLP-2). Texas, USA. pp: 18-24. Available at: http://delivery.acm.org/10.1145/980000/974239/p18-rich.pdf?key1=974239&key2=1050256621&coll=GUIDE&dl=GUIDE&CFID=78494647&CFTOKEN=91552183

Sánchez León, F.; Nieto Serrano, A. 1995. Development of a Spanish Version of the Xerox Tagger. Technical Report. Madrid: *Universidad Autónoma de Madrid*. Available at: http://arxiv.org/PS_cache/cmp-lg/pdf/9505/9505035v1.pdf

Santos, D. 2008. Resolução de anáfora pronominal em português utilizando o algorítimo de Hobbs. Master dissertation, Universidade Estadual de Campinas. Campinas, São Paulo. Available at: http://libdigi.unicamp.br/document/?code=000431264

Sasano,R.; Kawahara, D.; Kurohashi, S. 2008. A Fully-Lexicalized Probabilistic Model for Japanese Zero Anaphora Resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING2008)*. Manchester, United Kingdom. pp.769-776. Available at: http://aclweb.org/anthology/C/C08/C08-1097.pdf

Seki, K; Fujii, A.; Ishikawa, T. 2001. A Probabilistic Model for Japanese Zero Pronoun Resolution Integrating Syntactic and Semantic Features. *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium* (NLPRS2001). pp.403-410. Available at: http://www.cl.cs.titech.ac.jp/~fujii/paper/nlprs2001.pdf

Sidner, C. 1979. *Toward a computational theory of definite anaphora comprehension in English*. Technical report Nº. AI-TR-537. Cambridge, Massachussetts: MIT Press.

Soon, W; Ng, H.; Lim, C. 1999. Corpus-based learning for noun phrase coreference resolution. *Proceedings of the 1999 joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*. University of Maryland, USA. pp: 285-291. Available at: http://acl.ldc.upenn.edu/W/W99/W99-0634.pdf

Soon, W.; Ng, H.; Lim, C. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4). pp: 521-544. Available at: http://www.aclweb.org/anthology/J/J01/J01-4004.pdf

Trouilleux, F. 2002. A rule-based pronoun resolution system for French. *Proceedings of the Fourth Discourse Anaphora and Anaphora Resolution Colloquium (DAARC'02)*. Libon, Portugal. Available at: http://hal.archives-ouvertes.fr/docs/00/37/33/30/PDF/ftrouilleux_DAARC2002.pdf

Uehara, S. 1996. Anaphoric pronouns in English and their counterparts in Japanese. *Proceedings of the Discourse, Anaphora and Reference Resolution Conference* (DAARC96). Lancaster, UK. pp 64-75.

Vieira, R.; Poesio, M. 2000b. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4). pp: 525-579. Available at: http://www.aclweb.org/anthology-new/J/J00/J00-4003.pdf

Yeh, C.L.; Chen, Y.C. 2003. Zero anaphora resolution in chinese with partial parsing based on centering theory. *Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on*. pp: 683 – 688. Available at: http://www.colips.org/journal/volume17/JCLC_2007_V17_N1_04.pdf

Yeh, Ching-Long; Chen, Yi-Chun. 2007. Zero Anaphora Resolution in Chinese with Shallow. *Journal of Chinese Language and Computing 17 (1): 41-56*. Available at: http://www.colips.org/journal/volume17/JCLC_2007_V17_N1_04.pdf

Appendix

Appendix 1 – List of conjunctions

Coordinate sentence	Common conjunctions	
Additive	*e*, *nem* 'and', 'nor' or 'neither'	
Adversative	*mas* 'but'	
Alternative	*ou* 'or'	
Subordinate sentence		
Nouninal clause		
Type	Common conjunctions	
Integrant conjunction	*que*, *se* 'that', 'whether'	
Adverbial clause		
Conditional	Finite clause	Non-finite clause
	a não ser que, caso, *desde que*, *se*, *sem que*, *uma vez que*	*a*, *no caso de, na condição de*
	'if', 'unless'	
Causal	*como, dado que, já que*, *pois*, *porque*, *uma vez que*, *visto que*	*por*, *por causa de*, *devido ao fato de*
	'since', 'because'	
Finality	*a fim de que, para que*, *para*, *que*	*para*, *a fim de*
	'in order to', 'so that', 'in order that'	
Concessive	*embora*, *ainda que*, *posto que*, (*se*) *bem que, mesmo que*	*apesar de*
	'although', 'though', 'while'	
Time	*agora que, antes que, assim que, até que, depois que, desde que, enquanto, logo que, quando, sempre que*	*antes de, depois de*
	'when', 'before', 'after', 'since', 'while', 'as', 'as long as', 'until'	
Consecutive	*de forma que, de maneira que*, *de modo que*, *de sorte que*, *que* (preceded by the words - *tal*, *tanto*, *tão* or *tamanho*), *que*	
	'so… that'	

Appendix 2 – Annotation Guidelines

Presentation

A corpus with annotated zero anaphors has been created for the development of an anaphora resolution system for Portuguese. This document describes the annotation guidelines followed in the creation of this corpus. At this stage the corpus only contains texts from Brazilian Portuguese.

General notation

Zero anaphors are marked by a zero symbol *0* inside brackets *[]*, followed by an equal sign = and the arrow symbols < and >, corresponding to anaphora (1) and cataphora (2) relations, respectively, and a word indicating the head of the antecedent noun phrase (NP).

(1) *Um forte terremoto (6 graus na escala Richter) sacudiu ontem Taiwan,* **[0=< terremoto]** *provocando uma morte e ferimentos em duas pessoas*

'A strong earthquake (measuring 6 degrees on the Richter scale) shook Taiwan yesterday; the earthquake caused one death and injured two people'

(2) *Ao* **[0=>descobertas]** *apontarem para a cura de doenças atacando-as na escala infinitesimal dos genes, as novas descobertas da ciência representam um novo marco na linha de pensamento iniciada no século XIX pelo naturalista inglês Charles Darwin, autor da teoria da evolução*

'Pointing to the cure of diseases by attacking them in the infinitesimal scale of genes, the new discoveries of science represent a new milestone in the line of thought that has been started in the nineteenth century by the English naturalist Charles Darwin, author of the theory of evolution'

Only deleted subject of non-auxiliary verbs are to be market (3):

(3) *Essas células viajariam pelo corpo até os órgãos sexuais e de lá* **[0=<células]** *passariam às gerações seguintes*

'These cells would travel throughout the body until reaching sexual organs and from there they pass to next generations'

Verbal chains with auxiliary verbs whose subject has been zeroed count as a single verb form, hence there will be only one anaphor marked (4):

(4) *Mais de 90% dos machos descendentes das cobaias apresentavam os mesmos problemas, sem nunca* **[0=<machos] terem sido expostos** *ao inseticida*

'Over 90% of male descendants of the [experiment] subjects showed the same problems without ever having been exposed to insecticide'

The zeroed subject of non-finite dependent clauses is usually to be marked (5):

(5) *Do estudo resultou um mapa com a posição de cada uma das múltiplas variações dos genes, os tijolos moleculares que se combinam no coração das células para* **[0=<tijolos]** *definir as características físicas dos seres humanos*

'From the study, [it] resulted a map with the position of each one of the multiple variations of the genes, the molecular building blocks that combine themselves in the heart of cells to define the physical characteristics of humans beings'

In coordinated clauses only the zeroed subject of explicit verb forms is marked (6):

(6) *O profeta o obsedia e* **[0=<profeta]** *o persegue tanto que* **[0=<profeta]** *o vê em todo lugar;* **[0=<<profeta]** *preenche literalmente a paisagem, o que torna a ilusão visual…*

'The prophet obsesses him and [he=the prophet] pursues him so much that he sees him everywhere; [the prophet] literally fills the landscape, which makes the visual illusion…'

If the zeroed subject refers to a subordinate clause, then the anaphor will be noted [0(clause)=X] where *X* indicates the main verb of the antecedent clause (7):

(7) *"Esconder um programa desta magnitude não é apenas inapropriado, mas* **[0(clause)=esconder]** *é também ilegal", disse o senador democrata Dick Durbin*

'"Hiding a program of this magnitude is not only inappropriate but [it] is also illegal," said democratic senator Dick Durbin'

However, in some sentences, the reduced material cannot be easily recovered from the preceding discourse, hence, even if the anaphor type may be indicated, the antecedent proper is left unknown *?* (8):

(8) *Como não* [0=1p] *estamos vendo nossos espectadores,* [0=1p] *somos incapazes de* [0=1p] *observar sua reação ao que* [0=1p] *estamos fazendo e, com isso,* [0=1p] *ficamos à vontade para* [0=1p] *nos expor mais do que* **[0(clause)=?]** *seria prudente*

'Since [we] are not seeing our viewers, [we] are unable to observe their reaction to what we are doing and so [we] were at ease to expose ourselves more than [it] would be prudent [to do]'

On coordinated relative clauses, where the second relative pronoun has been zeroed (9), it should be marked but with the special notation [0(que)=<X], where *X* represents the antecedent of the relative pronoun:

(9) *Os processos epigenéticos também podem ocorrer pela modificação das histonas, as linhas que envolvem o DNA e* **[0(que)=<linhas]** *formam um novelo*

'The epigenetic process can also occur by the modification of histones, the lines that involve the DNA and form a ball'

In the example (9) above, the zero anaphor is placed *after* the coordinative conjunction *e* 'and'. In coordinated relative clauses with conjunction *nem* 'nor' (10), the zero anaphor is also placed *after* the conjunction, even if this representation may not be completely adequate:

(10) *"Não tenho nada: café, açúcar, nada", enfatiza o homem, que nunca se casou nem* **[0(que)=<homem]** *teve filhos, e* **[0(que)=<homem]** *não sabe ler*

'"[I] have nothing: coffee, sugar, nothing", emphasizes the man, who never married and [who] had no children and [who] cannot read'

Noun phrases

NP head nouns

For NPs whose head is a nominal determiner, for example *conjunto* 'set' (11), it is this head noun that the zeroed anaphor is referred to, even if the semantic head of the noun phrase is the complement of that determiner:

(11) *O terceiro fenômeno epigenético consiste na ação dos micro-RNAs, um conjunto de nucleotídeos que percorre o genoma* **[0=<conjunto]** *ligando e* **[0=<conjunto]** *desligando os genes*

'The third epigenetic phenomenon consists in the action of micro-RNAs, a set of nucleotides that travel the genome connecting and disconnecting the genes'

Numeral-nominal determinants such as *milhão* 'million', *milhar* 'thousand', linked to the determined noun by preposition de 'of' are not taken in consideration as antecedent of zero anaphors; instead, the head noun is the *N* they determine (12):

(12) *Segundo a última contagem do IBGE,* **23,5 milhões de pessoas** *vivem na Amazônia.* **[0=<<pessoas]** *São apenas 13% da população brasileira, mas o suficiente para* [0=<o] *fazer um estrago de proporções planetárias*

'According to the last count of IBGE, 23.5 million people live in the Amazon. [They] are only 13% of the Brazilian population, but enough to produce damage of planetary proportions'

In the case of adverbial determinants, such as *por cento* or *%* 'percent' (13), the antecedent of the zero anaphor is the head noun determined by these expressions:

(13) **Mais de 90% dos machos descendentes das cobaias** *apresentavam os mesmos problemas, sem nunca* **[0=<machos]** *terem sido expostos ao insecticida*

'Over 90% of male descendants of the [experiment] subjects showed the same problems without ever having been exposed to insecticide'

If the head noun has been zeroed in front of nominal determiners, the determinative noun is then taken as the head noun of the *NP* and may then function as antecedent for a zero anaphor (14):

(14) *Já as garotas tiveram resultados melhores: 75% dos homens toparam no ato. Dos 25% restantes,* **a maioria** *pediu desculpas,* **[0=<maioria]** *explicando que* **[0=<maioria]** *tinha marcado de* **[0=<maioria]** *sair com a namorada*

'On the other hand the girls had better results: 75% of men immediately agreed. From the remaining25%, the majority apologized, explaining that [they] already had a date with their girlfriend'

The noun *cento* 'cent' in the adverbial *por cento* 'percent' and the symbol *%* can be the antecedent of zeroed *NP* (15):

(15) *Quase todos os estudantes passaram de ano.* **Só 25%** *teve notas inferiores a 5,0 e* **[0=<%]** *tiveram de fazer recuperação*

'Nearly all students were approved. Only 25% had less than 5.0 and [they] had to do the retrieval'

Compound nouns

In the case of compound nouns, only the head noun is to be referred to in the zeroed anaphor (16):

(16) *Para* **[0=>Ministério]** *tentar incentivar a criação de mais mestrados profissionais no País, o* **Ministério da Educação** *publica hoje uma portaria [0=<portaria] estabelecendo novas regras para o credenciamento e a avaliação desses cursos*

'In order to try encouraging the creation of more professional master courses in the country, the Ministry of Education publishes today an ordinance establishing new rules for accreditation and evaluation of these courses'

Because of tokenization criteria, prefixed nouns are considered a compound word (e.g. *ex-colegas* 'ex-partners') (17):

(17) *Um exemplo conhecido dos adeptos do Orkut no Brasil são os* **ex-colegas** *de escola que, depois de anos sem* **[0=<ex-colegas]** *se comunicar e mesmo sem* **[0=<ex-colegas]** *ter nenhuma afinidade pessoal,* **[0=<ex-colegas]** *passam a engordar a lista de amigos virtuais uns dos outros*

'A known example of Orkut supporters in Brazil are the ex-school mates who, after years without communicating, even without having any personal affinity, start engrossing the list of each other's virtual friends'

Compound pronoun *a gente*, corresponding to a first person plural 'we', but imposing a third singular verbal agreement, will be referred to by the form *gente* (18):

(18) — *Mas* **a gente** *queria* **[0=<gente]** *ver filme, não show*

'— But we wanted to see a film, not a show'

The same happens with indefinite pronoun *todo (o) mundo* 'everyone' (19), which will be referred to by the head noun *mundo*:

(19) *E nem* **todo mundo** *aprendeu a* **[0=<mundo]** *usá-los a seu próprio favor*

'And not everyone learned how to use them to their own advantage'

Other compound (frozen) expressions, syntactically non-analisable are left without notation (20):

(20) *[…] genes […]. São eles que ensinam aos outros genes o* **caminho a seguir**, *para* [0=<eles] *dar continuidade às espécies [...]*

'[…] genes […]. It is them that teach others genes the way forward, in order to give continuity to the species'

Other half-frozen expressions with infinitive verbs are not marked (21):

(21) **No decorrer das décadas**, *no entanto, a população acabou se aprofundando na miséria*

'Over the decades, however, people just went deeper into poverty'

In this example, the subject of the nominalised verb *decorrer* is not marked.

Named entities

Compound proper names (named entities, in majuscules) are considered a single token and therefore, will be referred to in the notation of zero anaphors (22):

(22) *Lev Grossman, colunista da revista..., revelou* [0=impers] *há pouco* **[0=<Lev Grossman]** *ter decidido* **[0=<Lev Grossman]** *cancelar sua conta no Twitter [...]*

'Lev Grossman, columnist of the magazine …, revealed recently that [he] had decided to cancel his Twitter account […]'

In the case of titles in apposition with proper names, the two elements are considered the head noun of that NP (23)-(24):

(23) *No artigo que* **[0=>presidente Luiz Inácio Lula da Silva]** *escreveu especialmente para esta edição,* **o presidente Luiz Inácio Lula da Silva** *diz que "as soluções para a Amazônia têm de ser maiores que governos e mandatos, [0=<soluções] têm de ser assumidas pela sociedade brasileira e suas instituições"*

'In the article that [president Luiz Inácio Lula da Silva] wrote especially for this edition, president Luiz Inácio Lula da Silva says that "the solutions for the Amazon region must be larger than governments and mandates, [the solutions] must be undertaken by the Brazilian society and its institutions'

(24) **Dona Marta** *ficou um pouco preocupada com a chuva, pois [0=impers] haviam algumas falhas no telhado.* **[0=<<Dona Marta]** *Passou mais algumas recomendações tão peculiares às mães e* **[0=<<Dona Marta]** *encerrou a ligação*

'Mrs. Marta was a little worried about the rain, for there were some holes in the roof. [She] gave a few more recommendations, so typical of mothers, and hang up [the phone] connection'

Coordinated antecedent NPs or PPs

In the case of coordinated antecedent NPs or PPs, only the first head noun is to be referred to by the zero anaphor, but with the special notation *&* after that head noun (25):

(25) *De acordo com comunicado da Abia, ácaros e insetos estão presentes nas frutas e* **[0=<ácaros, &]** *se fragmentam quando* **[0=<ácaros, &]** *passam por máquinas processadoras de alimentos*

'According to the press release from Abia, mites and insects are present in fruits and [they] become fragmented when they pass through food processing machinery'

Pronominal use of articles and demonstratives

With the so-called pronominal use of definite and indefinite articles, as well as with demonstrative pronouns, the zeroed noun is *not* to be referred to in the following zero anaphor and hence a pronominal analysis is adopted for these words (26):

(26) *E os demais, apesar de* **[0=<os]** *serem titulados, terão de ter experiência profissional na área do curso*

'And the remaining [students], although [they] have already graduate, will have to acquire professional experience in the course's area'

Indefinite subject

General case

The indefinite subject is annotated as **[0=indef]** (27)-(28):

(27) **[0=indef]** *Nascer com patrimônio genético idêntico não significa que as pessoas crescerão tendo corpo, mente e doenças iguais*

'To be born with identical genetic heritage does not mean that people will grow up with similar body, mind and disease'

(28) *Apesar de todos os avanços na ciência da genética, apenas dentro de uma ou duas décadas será possível* **[0=indef]** *prevenir o aparecimento de doenças* **[0=indef]** *auscultando os genes, ou* **[0=indef]** *produzir remédios personalizados que ajam sobre o genoma específico de um paciente*

'Despite all the advances in genetic science, only in one or two decades will it be possible to prevent diseases from appearance by checking the genes, or to produce personalized medicines acting on the specific genome of a patient'

In case of (syntactically justified) doubt between indefinite anaphor and an antecedent NP, the indefinite anaphor is chosen (29):

(29) *Apesar do êxito de experiências pontuais para* **[0=indef]** *alterar o comportamento dos genes por meio de mudanças na alimentação [...]*

'Despite the success of occasional experiments to change the behavior of genes through diet changes'

In this case, the zeroed subject of *alterar* 'to change' could also refer to *experiências* 'experiments'.

In coordinated clauses, the zero anaphor is marked **[0=indef]** as usual (30):

(30) *Os cientistas estão ainda engatinhando no conhecimento de como* **[0=indef]** *ligar e* **[0=indef]** *desligar os genes*

'Scientist are still crawling in the knowledge of how turn on and turn off the genes'

Indefinite first person plural subject (1p)

Indefinite elliptical subject where there is a systematic ambiguity with first person plural *nós* 'we', will be specially noted **[0=1p]** (31):

(31) *As descobertas são impressionantes.* **[0=1p]** *Conseguimos informações preciosas sobre os genes, as marcas epigenéticas e as mudanças do genoma ao longo da vida, o que dá início a uma revolução*

'The findings are impressive. We got valuable information about the genes, the epigenetic markings and the changes of the genome throughout life, which initiates a revolution'

In this example, the first person plural may correspond to: a) a real plural, referring to the speaker and his/her team of researchers; b) a modesty plural, referring to the speaker; or c) the indefinite (generic) subject, referring to the scientific community as a whole. Naturally, such ambiguities cannot be solved at this stage.

Indefinite third person plural subject (3p)

Sentences with zeroed subject and the verb in the third person plural will be annotated **[0=3p]** (32)-(33); this type of subject is systematically ambiguous between the indefinite subject and are simple third person plural, and only context can disambiguate it:

(32) *Estou esperando o que me* **[0=3p]** *garantiram [...]*

'[I] am waiting what [they] assured me'

(33) *"Ainda* **[0=3p]** *estão fazendo isso lá embaixo",* [0=<<Zé Lopes] *acrescenta, sobre as praias sem vigilância ao longo do Rio Jutaí, um afluente do Solimões*

'"[They] are still doing it down there," [Zé Lopes] adds, about the beaches without surveillance along the Jutaí river, a tributary of the Solimões'

In case the antecedent of a zero anaphor cannot be precisely determined, a question mark will be used instead **[0=?]** (34):

(34) O encontro acontecera de repente, mas **[0=?]** era como se [0=3p] já tivessem sido amigos a vida inteira

'The meeting happened suddenly, but [it] was as if [they] has been friends for [their] entire life'

Note: This is a last resort solution, and should be used sparingly.

Impersonal subject

The impersonal subject is annotated as **[0=impers]**. This notation may cover different syntactic and semantic structures, such as meteorological constructions (35):

(35) — Nossa. **[0=impers]** Esfriou!

'— Wow. It got cold!'

and impersonal constructions with *haver* 'to there be' (36):

(36) "**[0=impers] Há** *uma perigosa tendência a* [0=indef] *fazer correlações entre etnia, crime e predisposição genética", alerta Pamela Sankar, professora de bioética da Universidade da Pensilvânia.*

'"There is a dangerous tendency to establish correlations between ethnic origin, crime and genetic predisposition", alerts Pamela Sankar, Bioethics professor at Pennsylvania University.'

or other impersonal verbs like *tratar-se de* 'to concern' 'to regard' (37):

(37) *Normalmente, faz-se referência à Amazônia Legal quando* **[0=impers] se trata** *de dados econômicos; as estatísticas sobre desmatamento – ou desflorestamento – dizem respeito apenas às áreas de floresta*

'Usually one uses the term Legal Amazon when it refers to economic data; the statistics on deforestation concern only the forest areas'

or temporal expressions (38):

(38) **[0=impers] Há muito [tempo]** *os cientistas sabem que o ambiente uterino atua de modo a* [0=<ambiente] *evitar que as informações genéticas embaralhadas dentro do zigoto produzam seres monstruosos*

'A long time ago, scientists know that the uterine environment acts in such a way so that [it] avoids that scrambled genetic information inside the zygote produced may monstrous beings'

Impersonal constructions with verbs *ter* (in Brazilian Portuguese) and *haver* (both in Brazilian and European Portuguese) may appear with a NP and a gerund (BP/EP) (39)-(40) or a prepositional infinitive (41) (only in EP):

(39) **[0=impers]** *Tem gente* **fazendo** *isso*

'There is people doing this'

(40) **[0=impers]** *Há gente* **fazendo** *isso*

'There is people doing this'

(41) **[0=impers]** *Há gente* **a fazer** *isso*

'There is people doing this'

In spite of the superficially complex structure of these sentences, we consider that the NP is the subject of the gerund (or of the prepositional infinitive (41)), so that no real reduction has effectively taken place and, therefore, no zero anaphor is to be marked.

Coreference chains

If in a coreference chain there are several coreferent NPs that can function as antecedent to a zero anaphor, the syntactically immediate antecedent is chosen (42):

(42) *A grande questão, ele completa, não é como* **as crianças** *aprendem a* **[0=<crianças]** *agredir, mas como* **elas** *aprendem a* **[0=<elas]** *não fazer isso*

'The big question, he adds, is not how the children learn to be aggressive, but how [they] learn not to do that'

Also in a coreference chain, when the antecedent of a zero anaphor is in a previous sentence[49], the notation [0=<<X] is used (43):

(43) **Os participantes** *concordaram com um programa ousado de combate à deterioração da terra, do ar e da água. Também* **[0=<<participantes]** *decidiram* **[0=<<participantes]** *buscar o crescimento econômico sem* **[0=<<participantes]** *degradar o meio ambiente*

'The participants agreed on a bold program for combating the deterioration of land, air and water. [They] also decided to pursue economic growth without degrading the environment'

even if the first element is in a fronted subordinate clause (44):

(44) *[...] Eco 92 [...]. Se* **[0=<<Eco 92]** *fracassar,* **[0=<<Eco 92]** *apagará a esperança de* **[0=<<Eco 92]** *dotar a comunidade internacional de uma tábua de mandamentos práticos e morais capaz de* [0=<tábua] *substituir o vácuo das ideologias*

'[…] Eco 92 […]. If Eco 92 fails, [it] will erase the hope of providing the international community of a board of practical and moral commandments able to replace the ideological vacuum '

The zero anaphor will be marked [0=<<X], no matter how many sentences away it may be. However, if in the discourse the first person plural is used as an indefinite and there is no necessary coreference chain between two (far apart) instances, the antecedent < or > sign is not used.

[49] The separators ';' and ':' are considered sentence boundaries, along with the common sentence separators ('.', '?', '!', etc.).

Coreference chains involving zero anaphors

In a coreference chain within the same sentence, if the antecedent of a zero anaphor *0_2* is also another zero anaphor *0_1*, the head of the antecedent NP of the latter *0_1* is repeated (45):

(45) *Ela ajudará na criação de* **remédios** *personalizados, capazes de* **[0=<remédios]** *alterar o genoma para* **[0=<remédios]** *deter o desenvolvimento de doenças e de transtornos psíquicos*

'[It] will help in the creation of personalized medicine, capable of altering the genome in order to stop the development of diseases and mental disorders'

This does not imply that the *0_2* refers **directly** to the antecedent of 0_1 (first occurrence); in the example (45), the reduction of the subject of the final subordinate clause *para 0_2 deter o desenvolvimento*… is not directly dependent of the antecedent head noun *remédios*. The analysis of the coreference chain is thus simplified.

In certain cases, a coreference chain can be determined among indefinite subjects; in this situation, the coreference relation is marked **[0=<indef]** if the zeroed element is in a subordinate clause (46):

(46) *A lista de amigos virtuais é uma espécie de agenda de telefones, com a vantagem de não ser necessário* [0=indef] *ligar para todos uma vez por ano para* **[0=<indef]** *não ser esquecido*

'The list of virtual friends is a kind of phonebook, with the advantage of not being required that one should call everyone once a year in order not to be forgotten'

The same happens with other indefinite subjects, such as the first person plural (1p) (47), and the third person plural (3p) (48):

(47) *Durante três meses* [0=1p] *percorremos a Amazônia para* **[0=<1p]** *revelar as tragédias e* **[0=<1p]** *conhecer as experiências que poderão preservar a mais rica biodiversidade da Terra*

'During three months, [we] traveled through the Amazon region to reveal the tragedies and to known the experiments that may preserve to the Earth's most rich biodiversity'

(48) *"*[0=3p] *Falaram que* [0=?] *ia trazer melhoria,* [0=?] *não trouxe nada", disse José Vasconcelos de Lima, o Zé Cigano. "Até* [0=indef] *comer está difícil no Paraíso.* **[0=<3p]** *Não consentiam a gente pescar,* [0=gente] *pegar tracajá."*

'"[They] said that [something?] would bring an improvement, [but] [it?] did not bring anything", said José Vasconcelos de Lima, the Gipsy Zé. "Even eating is hard in Paradise. [They] do not consent us to fish, [or] to take tracajá [fresh water turtle]."'

Exclusions

Adjectives

The subject of adjectives is only marked if they appear with their copula verb (e.g. *ser*, *estar*, 'to be') (49):

(49) *O mundo científico ficou ainda mais complexo depois do mapeamento genético feito há seis anos, quando os pesquisadores passaram a se dedicar a entender a função de cada um dos genes e, o supremo desafio,* [0=<pesquisadores] *explicar as razões pelas quais eles às vezes exercem suas funções e outras* [0=<eles] *parecem hibernar preguiçosamente nos cromossomos sem nunca* **[0=<eles]** *ser* <sic> *ativados* […]

'The scientific world became even more complex after the genetic mapping made six years ago, when the researchers began to devote themselves to the understanding of the function of each gene and, the ultimate challenge, to explain the reasons why they sometimes perform their functions and other times they seem to hibernate lazily in the chromosomes without ever being activated'

Therefore the zeroed subjects of adjectives in apposition are not marked (50):

(50) *Ela ajudará na criação de remédios personalizados,* **capazes** *de* [0=<remédios] *alterar o genoma para* [0=<remédios] *deter o desenvolvimento de doenças e de transtornos psíquicos*

'It will help in the creation of personalized medicine, capable of altering the genome in order to halt the development of diseases and mental disorders'

Past participles

The past participle is considered as an ordinary adjective and its zeroed subject should be marked accordingly depending on the presence ((51)-(52)) or absence (53) of the copula verb:

(51) *Certamente* **[0=<marido]** *estava armado*

'Certainly the husband was armed'

(52) *Darwin sentiu o peso, e* **[0=<Darwin]** *ficou aterrorizado*

'Darwin felt the weight, and [he] was terrified'

(53) *Hoje, líderes indígenas* **formados** *em universidades dirigem entidades e [0=<líderes] se espelham em Evo Morales, o índio aimará que preside a Bolívia.* (no mark-up)

'Today, indigenous leaders trained in universities lead several institutions and [feel that they] are mirrored in Evo Morales, the Aymara Indian who presides over Bolivia'

The past participle is considered a verbal form when it makes part of a compound tense with auxiliary verbs ter (54) 'to have' or (rarely) haver 'to there be' (55):

(54) *"Eles precisam de tempo e de intimidade; como diz o ditado,* [0=<eles] *não podem se conhecer sem que* **[0=<eles] tenham comido** *juntos a quantidade necessária de sal"*

'"They need time and intimacy; as the saying goes, [they] cannot cannot know each other without having eaten together the necessary quantity of salt"'

(55) *Apesar de* **[0=>Arthur] haver errado** *todos os seis tiros, Artur conseguiu afastar a criatura.* [0=<Arthur] *Ajudou o senhor José a levantar*

'Although Arthur had failed all six shots round, he managed to keep the creature away. [He] helped Mr. José to stand up'

Reduced gerundives

Like adnominal and appositional adjectives, in reduced gerundives resulting from relative clauses the subject is considered to be explicit and it is not marked (56)-(57):

(56) *Luiz percebeu faíscas* **saindo** *de um poste à frente da casa*

'Luiz saw sparks coming out of a pole in front of the house'

(57) *=Luiz percebeu faíscas* **que estavam saindo** *de um poste à frente da casa*

='Luiz saw sparks that were coming out of a pole in front of the house'

Otherwise gerundive adverbial clauses need the marking of zeroed subjects (58):

(58) *Essas mudanças podem ser para o bem ou para o mal,* **[0=<mudanças] atenuando** *sintomas de doenças ou* **[0=<mudanças] provocando** *seu desenvolvimento*

'These changes can be for good or for evil, alleviating symptoms of disease or causing their development'

Topicalization structures and other forms of focus

Topicalization structures and other forms of focusing sentence elements involving changes in sentences' basic word-order are not marked and the syntactic position left empty by the moved constituent (59) is not signaled:

(59) De fato pesava bastante, o tal saco

'Indeed [it] weighed a lot, that bag'

In much the same way, cleft sentences with *ser … que* are not marked for their subject NPs (60):

(60) *É nas trilhas desse vazio,* [0=>aventureiros] *desfraldando falsas bandeiras do progresso,* **que** *aventureiros nacionais e internacionais invadiram a floresta e* [0=<aventureiros] *desataram as tragédias*

'It is in the trails of this gap, unfurling the false flags of progress, that the national and international adventurers have invaded the forest and have untied the tragedies'

Imperative, interrogative and exclamative sentences

The zeroed subject of imperative sentences (61); direct, total (yes/no) (62) or partial (*wh-*) (63)-(62); interrogative sentences; question tags (65); and exclamative sentences (64) are not to be marked:

(61) Saia um pouco da sua página virtual, pare de bisbilhotar a dos outros, dê um tempo nas conversinhas que só pontuam o vazio da existência e vá viver mais.

'Get out of your virtual page, stop snooping around the pages of other people, take a timeout from those chats that only punctuate the emptiness of one's existence and do have life of your own'

(62) Não ouviu falar?

'Did [you] ear [someone] saying [anything about it] ?'

(63) O que está esperando?

'What are [you] waiting for?'

(64) — E abrir a janela? Nem pensar! — Protestou Marina.

'— And [me] open the window? No way! — Protests Marina.'

(65) Amanhã você vai ficar em casa, não vai?

'Tomorrow [you] will be at home, wont you?'

For indirect interrogative subordinate clauses with interrogative *qu-* (*wh-*) pronouns (*question cachée*), the pronoun is considered the head of the clause and can be referred to by zero anaphor (66):

(66) **Quem não retribuir a oferta quando a situação for inversa** *fica com a reputação manchada e* **[0=<quem]** *é banido do almoço grátis*

'[He] who does not return the offer when the situation will be reversed, will have his reputation tarnished and will be banned from the free lunch'

Causative operator-verbs

On constructions of causative operator verbs (*Vopc*) with restructured subject, the structurally zeroed slot of the subject of the dependent clause is not marked (67)-(68):

(67) *A falta de comunicação com o resto da Terra* **permitiu ao regime permanecer** *mergulhado no passado.* (subject of *permanecer* is not marked)

(= A falta de comunicação com o resto da Terra permitiu [ao regime] que [o regime] permanecesse mergulhado no passado)

'The lack of communication with the rest of the globe has allowed to the regime to remain immersed into the past'

(68) *Que importava se num dia futuro sua marca* **ia fazê-la erguer** *insolente uma cabeça de mulher?*

'What does it matter if one future day her mark would make her rise outrageously a woman's head?'

Direct speech

In the case of direct speech (for example, in interviews) the first person subject and the second person (eventually the *você* personal pronoun, corresponding to a second person but imposing to the verb a third person agreement), if zeroed, are *not* to be marked (69):

(69) Quando fico conectada com um monte de gente por muito tempo, tenho a impressão de que, no fundo, não conheço ninguém

'When I stay connected with a lot of people for a long time, I have the impression that, basically, I do not know anyone'

Reduced, infinitive prepositional clauses

Reduced, infinitive prepositional clauses, usually resulting from the reduction of relative are treated as other relatives, that is, no zero anaphor is considered (70):

(70) *Os norte-coreanos não estão sendo tratados como os iraquianos porque avalia-se que a estratégia* **a ser seguida** *é [0=indef] impedir que um país inimigo consiga obter armas nucleares.*

'The North Koreans are not being treated as the Iraqis because it is assessed that the strategy [that is] being followed is to prohibit an enemy country from being able to obtain nuclear weapons'

In this example, the NP *a estratégia* **a ser seguida** 'the strategy being followed' is analyzed from the reduction of the relative clause *a estratégia que está sendo seguida* 'the strategy that is being followed'.

Appendix 3 – Set of written sentences

[1] Eu comi o bolo mas fiquei com fome.
[2] Tu comeste o bolo mas ficaste com fome.
[3] O Pedro comeu o bolo mas ficou com fome.
[4] A Joana comeu o bolo mas ficou com fome.
[5] Você comeu o bolo mas ficou com fome.
[6] Nós comemos o bolo mas ficámos com fome.
[7] Vós comestes o bolo mas ficastes com fome.
[8] Eles comeram o bolo mas ficaram com fome.
[9] Vocês comeram o bolo mas ficaram com fome.
[10] O Pedro acabou a corrida mas ficou cansado.
[11] A Joana acabou a corrida mas ficou cansada.
[12] Eles comeram o bolo mas ficaram cansados.
[13] Elas comeram o bolo mas ficaram cansadas.
[14] Vocês comeram o bolo mas ficaram cansados.
[15] Vocês comeram o bolo mas ficaram cansadas.
[16] O Pedro e a Joana comeram o bolo mas ficaram cansados.
[17] A Maria e a Joana comeram o bolo mas ficaram cansadas.
[18] Eu e a Joana comemos o bolo mas ficámos cansados.
[19] Eu e a Joana comemos o bolo mas ficámos cansadas.
[20] Tu e a Joana comestes o bolo mas ficastes cansados.
[21] Tu e a Joana comestes o bolo mas ficastes cansadas.
[22] Tu e a Joana comeram o bolo mas ficaram cansados.
[23] Tu e a Joana comeram o bolo mas ficaram cansadas.
[24] Comi o bolo mas fiquei com fome.
[25] Comeste o bolo mas ficaste com fome.
[26] Comeu o bolo mas ficou com fome.
[27] Comeu o bolo mas ficou com fome.
[28] Comeu o bolo mas ficou com fome.
[29] Comemos o bolo mas ficámos com fome.
[30] Comestes o bolo mas ficastes com fome.
[31] Comeram o bolo mas ficaram com fome.
[32] Acabou a corrida mas ficou cansado.
[33] Acabou a corrida mas ficou cansada.
[34] Comeram o bolo mas ficaram cansados.
[35] Comeram o bolo mas ficaram cansadas.
[36] Comeram o bolo mas ficaram cansados.
[37] Comeram o bolo mas ficaram cansadas.
[38] O Pedro foi ao cinema porque queria ver o filme.

[39] O Pedro foi ao cinema para ver o filme.
[40] O Pedro foi ao cinema a fim de ver o filme.
[41] Nós fomos ao cinema a fim de vermos o filme.
[42] Nós fomos ao cinema a fim de ver o filme.
[43] *Eu fui ao cinema a fim de veres o filme.
[44] Tu foste ao cinema a fim de veres o filme.
[45] *Ele foi ao cinema a fim de veres o filme.
[46] *Nós fomos ao cinema a fim de veres o filme.
[47] *Vós fostes ao cinema a fim de veres o filme.
[48] *Eles foram ao cinema a fim de veres o filme.
[49] *Eu fui ao cinema a fim de vermos o filme.
[50] *Tu foste ao cinema a fim de vermos o filme.
[51] *Ele foi ao cinema a fim de vermos o filme.
[52] Nós fomos ao cinema a fim de vermos o filme.
[53] *Vós fostes ao cinema a fim de vermos o filme.
[54] *Eles foram ao cinema a fim de vermos o filme.
[55] *Eu fui ao cinema a fim de assistirdes ao filme.
[56] *Tu foste ao cinema a fim de assistirdes ao filme.
[57] *Ele foi ao cinema a fim de assistirdes ao filme.
[58] *Nós fomos ao cinema a fim de assistirdes ao filme.
[59] Vós fostes ao cinema a fim de assistirdes ao filme.
[60] *Eles foram ao cinema a fim de assistirdes ao filme.
[61] *Eu fui ao cinema a fim de verem o filme.
[62] *Tu foste ao cinema a fim de verem o filme.
[63] *Ele foi ao cinema a fim de verem o filme.
[64] *Nós fomos ao cinema a fim de verem o filme.
[65] *Vós fostes ao cinema a fim de verem o filme.
[66] Eles foram ao cinema a fim de verem o filme.
[67] O Pedro foi ao cinema ver o filme.
[68] O Pedro estava farto de ver este filme.
[69] Eu estava farto de ver este filme.
[70] Tu estavas farto de ver este filme.
[71] Ele estava farto de ver este filme.
[72] Nós estávamos fartos de ver este filme.
[73] Vós estáveis fartos de ver este filme.
[74] Eles estavam fartos de ver este filme.
[75] *Eu estava farto de veres este filme.
[76] Tu estavas farto de veres este filme.
[77] *Ele estava farto de veres este filme.
[78] *Você estava farto de veres este filme.
[79] *Nós estávamos fartos de veres este filme.

[80] *Vós estáveis fartos de veres este filme.
[81] *Eles estavam fartos de veres este filme.
[82] *Vocês estavam fartos de veres este filme.
[83] *Eu estava farto de vermos este filme.
[84] *Tu estavas farto de vermos este filme.
[85] *Ele estava farto de vermos este filme.
[86] *Você estava farto de vermos este filme.
[87] Nós estávamos fartos de vermos este filme.
[88] *Vós estáveis fartos de vermos este filme.
[89] *Eles estavam fartos de vermos este filme.
[90] *Vocês estavam fartos de vermos este filme.
[91] *Eu estava farto de assistirdes a este filme.
[92] *Tu estavas farto de assistirdes a este filme.
[93] *Ele estava farto de assistirdes a este filme.
[94] *Você estava farto de assistirdes a este filme.
[95] *Nós estávamos fartos de assistirdes a este filme.
[96] Vós estáveis fartos de assistirdes a este filme.
[97] *Eles estavam fartos de assistirdes a este filme.
[98] *Vocês estavam fartos de assistirdes a este filme.
[99] *Eu estava farto de assistirem a este filme.
[100] *Tu estavas farto de assistirem a este filme.
[101] Ele estava farto de assistirem a este filme.
[102] Você estava farto de assistirem a este filme.
[103] *Nós estávamos fartos de assistirem a este filme.
[104] *Vós estáveis fartos de assistirem a este filme.
[105] Eles estavam fartos de assistirem a este filme.
[106] Vocês estavam fartos de assistirem a este filme.
[107] O Pedro leu o jornal palitando os dentes.
[108] O Pedro estava a ler o jornal palitando os dentes.
[109] O Pedro estava lendo o jornal palitando os dentes.
[110] O Pedro tinha lido o jornal palitando os dentes.
[111] O Pedro caiu palitando os dentes.
[112] O Pedro estava a cair palitando os dentes.
[113] O Pedro estava caindo palitando os dentes.
[114] O Pedro tinha caído palitando os dentes.
[115] O fumo cobre o céu compelindo os motoristas a acender os faróis.
[116] O fumo cobre o céu fazendo os motoristas a acender os faróis.
[117] O fumo cobre o céu forçando os motoristas a acender os faróis.
[118] O fumo cobre o céu levando os motoristas a acender os faróis.
[119] O fumo cobre o céu obrigando os motoristas a acender os faróis.
[120] O Pedro mandou a Ana lavar a loiça.

[121] O Pedro impediu a Ana de lavar a loiça.
[122] O fumo cobre o céu fazendo com que os motoristas acendam os faróis.
[123] O Pedro ordenou à Ana que lavasse a loiça.
[124] O Pedro exigiu à Ana que lavasse a loiça.
[125] O Pedro determinou que lavasse a loiça.
[126] O Pedro viu a Ana a ler o jornal.
[127] O Pedro encontrou a Ana a ler o jornal.

Appendix 4 – List of rules implemented

```
////////////////////////////////////////////////////////////////////////
////////////////////////////////////////////////////////////////////////
//
// ANAPHORA 0
//
////////////////////////////////////////////////////////////////////////
////////////////////////////////////////////////////////////////////////

//=======================================================
// ANAPHOR 0 SUBJECTS
//=======================================================

// Example: O rapaz comeu o bolo [mas]CONJ ficou com fome      ->
SUBJ_ANAPH0(ficou,rapaz)
//
|     #1[verb],     ?*,     CONJ[coord];PUNCT[lemma:";"];PUNCT[lemma:":"],
?*[verb:~,sc:~], #3[verb] |
 if  (  HEAD(#4,#1)  &  VDOMAIN(?,#4)  &  SUBJ(#4,#5)  &  HEAD(#6,#3)  &
VDOMAIN(#6,#7) & ~SUBJ(#7,?) &
      ((#5[person]:#7[person] & #5[number]:#7[number]) || (#5[person:~] &
#7[3p] & #5[number]:#7[number] & ~COORD(?,#5)) || (#5[person:~] & #7[3p,pl]
& COORD(?,#5)) || #7[person:~])
    )
   SUBJ[pre=+,anaph0=+](#7,#5)


//=======================================================

// Verbo principal s_pp_qufconj com SC
//  Example: O João ordenou à Ana que lavasse a loiça            ->
SUBJ_ANAPH0(lavar,Ana)
//
| #1[verb], ?*[verb:~], PP#8, SC{?*, ?#3[verb,last]} |
 if ( HEAD(#4[s_pp_qufconj],#1) & VDOMAIN(?,#4) & SUBJ(#4,#5) & HEAD(#6,#3)
& HEAD(#9,#8) & MOD[post](#4,#9) & VDOMAIN(#6,#7) & ~SUBJ(#7,?) )
   SUBJ[pre=+,anaph0=+](#7,#9)

// Caso geral com SC
//  Example: O João comeu batatas quando foi a Lisboa            ->
SUBJ_ANAPH0(foi,João)
//
| #1[verb], ?*[verb:~], SC{?*, ?#3[verb,last]} |
 if ( HEAD(#4[s_qufconj:~],#1) & VDOMAIN(?,#4) & SUBJ(#4,#5) & HEAD(#6,#3)
& VDOMAIN(#6,#7) & ~SUBJ(#7,?)   &
      ((#5[person]:#7[person] & #5[number]:#7[number]) || (#5[person:~] &
#7[3p] & #5[number]:#7[number] & ~COORD(?,#5)) || (#5[person:~] & #7[3p,pl]
& COORD(?,#5)) || #7[person:~])
    )
   SUBJ[pre=+,anaph0=+](#7,#5)




//=======================================================
```

```
// Example: O Pedro {foi}VF ao cinema SC{porque queria} VINF{ver} o filme -
> SUBJ_ANAPH0(queria, Pedro), SUBJ_ANAPH0(ver,Pedro)
// Example: O Pedro {foi}VF ao cinema VINF{para ver} o filme. ->
SUBJ_ANAPH0(ver,Pedro)
// Example: O Pedro {foi}VF ao cinema VINF{ver} o filme ->
SUBJ_ANAPH0(ver,Pedro)
// Example: O Pedro {estava}VCOP ADJP{farto} VINF{de ver} este filme ->
SUBJ_ANAPH0(ver,Pedro)
//
 if ( MOD[post,inf,sentential](#1,#7) & SUBJ[pre](#1,#5) & ~SUBJ(#7,?) )
   SUBJ[pre=+,anaph0=+](#7,#5)

// Example: O Pedro leu o jornal palitando os dentes ->
SUBJ_ANAPH0(palitando,Pedro)
//
 if ( MOD[post,gerund,sentential](#1,#7) & SUBJ[pre](#1,#5) & ~SUBJ(#7,?) )
   SUBJ[pre=+,anaph0=+](#7,#5)



//=========================================================

// Example: Quando o Pedro foi ao Porto, encontrou a Ana ->
SUBJ_ANAPH0(encontrou,Pedro)
//
|   ?*[verb],   SC{?*,   ?#1[verb,last]},   ?*[sc:~],   PUNCT[comma],
?*[verb:~,sc:~], ?#3[verb] |
 if ( HEAD(#2,#1) & SUBORD(?,#2) & VDOMAIN(#2,#4) & SUBJ(#4,#5) &
HEAD(#6,#3) & VDOMAIN(#6,#7) & ~SUBJ(#7,?) )
   SUBJ[pre=+,anaph0=+](#7,#5)


// Example: Quando veio ao Porto para comer bolos, o Pedro encontrou a Ana-
> SUBJ_ANAPH0(comer,Pedro)
//
| ?*[verb], SC{?*, ?#1[verb,last]}, ?*[sc:~], SC{?*, ?#3[verb,last]} |
 if ( HEAD(#2,#1) & SUBORD(?,#2) & VDOMAIN(#2,#4) & SUBJ(#4,#5) &
HEAD(#6,#3) & & SUBORD(?,#6) VDOMAIN(#6,#7) & ~SUBJ(#7,?) )
   SUBJ[pre=+,anaph0=+](#7,#5)


//=========================================================

// Example: O Pedro era esperto mas não era inteligente ->
ATTRIB_ANAPH0(Pedro,inteligente)
//
 if ( PREDSUBJ(#1[cop],#2) & SUBJ[anaph0](#1,#3) )
   ATTRIB[anaph0=+](#3,#2).

// Example: O Pedro era esperto mas não inteligente ->
ATTRIB_ANAPH0(Pedro,inteligente)
//
| #1[verb], ?*, CONJ[coord];PUNCT[lemma:";"];PUNCT[lemma:":"], (PP*;ADVP*),
AP#5 |
 if ( HEAD(#2,#1) & VDOMAIN(?,#2) & PREDSUBJ(#2,#3) & ATTRIB(#4,#3) &
HEAD(#6,#5) & ~ATTRIB(?,#6) )
   ATTRIB[anaph0=+](#4,#6)


//=== CATÁFORA ============================================
```

```
// Example: Quando veio ao Porto, o Pedro encontrou a Ana            ->
SUBJ_ANAPH0(veio,Pedro)
//
|    ?*[verb],    SC{?*,    ?#1[verb,last]},    ?*[sc:~],    PUNCT[comma],
?*[verb:~,sc:~], ?#3[verb] |
 if  (  HEAD(#2,#1)  &  SUBORD(?,#2)  &  VDOMAIN(#2,#4)  &  ~SUBJ(#4,?)  &
HEAD(#6,#3) & VDOMAIN(#6,#7) & SUBJ(#7,#5) )
   SUBJ[post=+,anaph0=+](#4,#5)


//=========================================================

// Example: A Joana e a Maria comeram o bolo mas ficaram com fome ->
SUBJ_ANAPH0(ficaram,Joana)
//                                                                   ->
SUBJ_ANAPH0(ficaram,Maria)
//
 if ( SUBJ[anaph0](#2,#1) && coord(#3,#1) & coord(#3,#4) && ~SUBJ(#2,#4) )
    SUBJ[anaph0=+,pre=+](#2,#4)

 if   (   SUBJ[anaph0](#2,#1)   &&   coord(#3,#1)   &   coord(#3,#4)   &&
^SUBJ[anaph0:~](#2,#4) )
    SUBJ[anaph0=+](#2,#4)
```

Appendix 5 – List of control verbs

```
//!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
//! L2F - Spoken Language Systems Laboratory / INESC-ID Lisboa           !
//!                                                                      !
//! Copyright (c) 2006 INESC-ID Lisboa.  All Rights Reserved.            !
//!                                                                      !
//! --- CONTENTS ---                                                     !
//!   Project    : Portuguese XIP grammar                                !
//!   filename   : lexVbControl.xip                                      !
//!   Date       : 2007/May                                              !
//!   Author     : Caroline Hagege                                       !
//!   DateRevier : 2010/May                                              !
//!   Reviewer   : Jorge Baptista & Simone Pereira                       !
//!                                                                      !
//!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
//!                                                                      !
//! Verbs that introduce completive subordinate clauses and that impose   !
//!constraints (control) ont the reference of SUBJ or OBJ of those clauses!
//!Some of these structures are derived from the transformation          !
//!(reduction) of completives, eventually the restructuring of some of    !
//!their arguments.                                                      !
//!                                                                      !
//!                                                                      !
//!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

///////////////////////////////////////////////////////////////////////////
///////////////////////////////////////////////////////////////////////////
//
// Feature description (and examples)
//
///////////////////////////////////////////////////////////////////////////
///////////////////////////////////////////////////////////////////////////

// TRAÇO s_inf
// significa que pode introduzir uma infinitiva cujo sujeito é tb sujeito
// do verbo
//
// Exemplo: O Pedro prometeu lavar a loiça

// TRAÇO s_np_inf
// significa que pode introduzir uma infinitiva cujo sujeito será o NP à
// esquerda da infinitiva
//
// Exemplo: O Pedro autorizou a Ana a lavar a loiça
// Exemplo: O Pedro impediu a Ana de lavar a loiça

// TRAÇO s_pp_inf
// significa que pode introduzir uma completiva *infinitiva* cujo sujeito
// será o mesmo que o complemento indirecto
//
// Exemplo: O Pedro pediu à Ana para lavar a loiça

// TRAÇO s_pp_qufconj
// significa que pode introduzir uma completiva  *finita* cujo sujeito
// será o mesmo que o do verbo assim marcado
//
// Exemplo: O Pedro ordenar à Ana que lavasse a loiça
```

```
// TRAÇO s_qufconj
// significa que pode introduzir uma completiva finita cujo sujeito não
// pode ser o mesmo que o do verbo principal
//
// Exemplo: O Pedro determinou que lavassem a loiça

// TRAÇO s_qufind (ainda não existe nenhum verbo com esta etiqueta)
// significa que pode introduzir uma completiva finita cujo sujeito é o
// mesmo que o do verbo principal
//
// Exemplo: O Pedro achou que estava cansado

////////////////////////////////////////////////////////////////////////
////////////////////////////////////////////////////////////////////////
//
// Lexicon
//
////////////////////////////////////////////////////////////////////////
////////////////////////////////////////////////////////////////////////

Vocabulary:

//permitir:       verb += [s_np_inf:+].   não admite uma infinitiva (só na
//forma reflexa - o reflexo é o sujeito da infinitiva)

// SÓ têm a propriedade:  s_np_inf
//    assumne-se que s_np_inf = o np é sujeito do inf

// SÓ têm a propriedade:  s_inf
//    assumne-se que s_inf    = o sujeito do inf é o sujeito da oração
//    principal

// Têm ambas as propriedades: s_np_inf e s_inf
//    assumne-se que s_np_inf = o np é sujeito do inf
//    assumne-se que s_inf    = o sujeito do inf é o sujeito da oração
//    principal

abominar:         verb += [s_inf:+,s_np_inf:+].
aceitar:          verb += [s_inf:+,s_np_inf:+].
achar:            verb += [s_np_inf:+].
acreditar:        verb += [s_inf:+].
adivinhar:        verb += [s_np_inf:+].
admitir:          verb += [s_inf:+,s_np_inf:+].
adorar:           verb += [s_inf:+].
aguardar:         verb += [s_inf:+,s_np_inf:+].
amaldiçoar:       verb += [s_inf:+,s_np_inf:+].
ambicionar:       verb += [s_inf:+].
anotar:           verb += [s_inf:+,s_np_inf:+].
ansiar:           verb += [s_inf:+,s_np_inf:+].
antever:          verb += [s_inf:+,s_np_inf:+].
apreciar:         verb += [s_inf:+,s_np_inf:+].
aprender:         verb += [s_inf:+].
aprovar:          verb += [s_inf:+,s_np_inf:+].
apurar:           verb += [s_inf:+].
argumentar:       verb += [s_inf:+].
assumir:          verb += [s_inf:+,s_np_inf:+].
autorizar:        verb += [s_np_inf:+,s_pp_inf:+].
averiguar:        verb += [s_inf:+,s_np_inf:+].
calcular:         verb += [s_inf:+].
cismar:           verb += [s_inf:+].
coagir:           verb += [s_np_inf:+].
```

```
compelir:         verb += [s_np_inf:+].
compreender:      verb += [s_inf:+,s_np_inf:+].
comprovar:        verb += [s_inf:+,s_np_inf:+].
conceber:         verb += [s_inf:+,s_np_inf:+].
conceder:         verb += [s_np_inf:+,s_pp_inf:+].
concluir:         verb += [s_inf:+,s_np_inf:+].
condenar:         verb += [s_np_inf:+].
conseguir:        verb += [s_inf:+].
considerar:       verb += [s_inf:+,s_np_inf:+].
constatar:        verb += [s_inf:+,s_np_inf:+].
contar:           verb += [s_inf:+].
costumar:         verb += [s_inf:+].                //Vaux
crer:             verb += [s_inf:+,s_np_inf:+].
cuidar:           verb += [s_inf:+,s_np_inf:+].
decidir:          verb += [s_inf:+].
//decretar:         verb += [s_infdif:+].
deduzir:          verb += [s_inf:+,s_np_inf:+].
defender:         verb += [s_inf:+,s_np_inf:+].
deixar:           verb += [s_inf:+,s_np_inf:+].
deliberar:        verb += [s_inf:+].
desanimar:        verb += [s_np_inf:+].
desaprovar:       verb += [s_qufconj:+,s_pp_qufconj:+].
descobrir:        verb += [s_inf:+,s_np_inf:+].
desconhecer:      verb += [s_inf:+,s_np_inf:+].
descortinar:      verb += [s_inf:+,s_np_inf:+].
desejar:          verb += [s_inf:+].
desencorajar:     verb += [s_np_inf:+].
desestimular:     verb += [s_np_inf:+].
desobrigar:       verb += [s_np_inf:+].
determinar:       verb += [s_inf:+,s_qufconj:+].
detestar:         verb += [s_inf:+].
dispensar:        verb += [s_np_inf:+].
encontrar:        verb += [s_np_inf:+].
encorajar:        verb += [s_np_inf:+].
entender:         verb += [s_inf:+,s_np_inf:+].
esperar:          verb += [s_inf:+,s_np_inf:+].
estabelecer:      verb += [s_inf:+].
estimar:          verb += [s_inf:+,s_np_inf:+].
estimular:        verb += [s_np_inf:+].
estipular:        verb += [s_inf:+,s_np_inf:+].
estranhar:        verb += [s_inf:+,s_np_inf:+].
exigir:           verb += [s_inf:+,s_np_inf:+,s_pp_inf:+,s_pp_qufconj:+,
                           s_qufconj].
exortar:          verb += [s_np_inf:+].
experimentar:     verb += [s_inf:+].
fazer:            verb += [s_np_inf:+].
fingir:           verb += [s_inf:+,s_np_inf:+].
forçar:           verb += [s_np_inf:+].
ignorar:          verb += [s_inf:+].
imaginar:         verb += [s_inf:+,s_np_inf:+].
impedir:          verb += [s_np_inf:+].
impelir:          verb += [s_np_inf:+].
incitar:          verb += [s_np_inf:+].
inibir:           verb += [s_np_inf:+].
insinuar:         verb += [s_inf:+,s_np_inf:+].
instigar:         verb += [s_np_inf:+].
inventar:         verb += [s_inf:+,s_np_inf:+].
isentar:          verb += [s_np_inf:+].
julgar:           verb += [s_inf:+,s_np_inf:+].
lamentar:         verb += [s_inf:+,s_np_inf:+].
lastimar:         verb += [s_inf:+,s_np_inf:+].
```

```
levar:            verb += [s_np_inf:+].
livrar:           verb += [s_np_inf:+].
lograr:           verb += [s_inf:+].
mandar:           verb += [s_np_inf:+,s_pp_inf:+].
manter:           verb += [s_np_inf:+].
merecer:          verb += [s_inf:+,s_np_inf:+].
notar:            verb += [s_inf:+,s_np_inf:+].
obrigar:          verb += [s_np_inf:+,s_pp_inf:+].
observar:         verb += [s_np_inf:+].
odiar:            verb += [s_inf:+].
opinar:           verb += [s_np_inf:+].
ordenar:          verb += [s_pp_qufconj:+].
parecer:          verb += [s_inf:+].        //Vcop
pedir:            verb += [s_pp_inf:+].
pensar:           verb += [s_inf:+].
perceber:         verb += [s_inf:+,s_np_inf:+].
permitir:         verb += [s_np_inf:+,s_pp_inf:+,s_pp_qufconj:+].
planear:          verb += [s_inf:+].
postular:         verb += [s_inf:+,s_np_inf:+].
precisar:         verb += [s_inf:+,s_qufconj:+].
pressentir:       verb += [s_inf:+,s_np_inf:+].
pressupor:        verb += [s_inf:+,s_np_inf:+].
presumir:         verb += [s_inf:+,s_np_inf:+].
pretender:        verb += [s_inf:+,s_qufconj:+].
procurar:         verb += [s_inf:+].
proibir:          verb += [s_np_inf:+,s_pp_inf:+].
prometer:         verb += [s_pp_inf:+].
querer:           verb += [s_inf:+,s_np_inf:+,s_qufconj:+].
ratificar:        verb += [s_np_inf:+,s_qufconj:+].
recear:           verb += [s_inf:+,s_np_inf:+].
reconhecer:       verb += [s_inf:+,s_np_inf:+].
reconsiderar:     verb += [s_inf:+].
resolver:         verb += [s_inf:+].
respeitar:        verb += [s_np_inf:+,s_qufconj:+].
saber:            verb += [s_inf:+].
sentir:           verb += [s_inf:+,s_np_inf:+].
simular:          verb += [s_inf:+].
sonhar:           verb += [s_inf:+,s_np_inf:+].
supor:            verb += [s_inf:+,s_np_inf:+].
suportar:         verb += [s_inf:+,s_np_inf:+].
temer:            verb += [s_inf:+,s_np_inf:+].
tencionar:        verb += [s_inf:+].
tentar:           verb += [s_inf:+].
topar:            verb += [s_inf:+,s_np_inf:+].
ver:              verb += [s_np_inf:+].
verificar:        verb += [s_inf:+,s_np_inf:+].
```