Linguistics Parameters for Zero Anaphora Resolution

Simone Pereira[1,2], Richard Evans[2], Jorge Baptista[1,3]
[1] Universidade do Algarve, Faculdade de Ciências Humanas e Sociais
[2] University of Wolverhampton, School of Law, Social Sciences and Communications
[3] L2F Spoken Language Laboratory - INESC ID Lisboa
simonecp@gmail.com, R.J.Evans@wlv.ac.uk, jbaptis@ualg.pt

Abstract. This paper 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. It describes the main grammatical rules imposing subject NP deletion and referential constraints in Brazilian Portuguese, in order to allow a correct identification of the antecedent of the deleted subject NP. These rules were then formalized into the Xerox Incremental Parser (XIP), rule-based, deep parsing as a module of Portuguese grammar developed at Spoken Language Laboratory (L2F). A corpus of different text genres was manually annotated for zero anaphors and other zero-shaped, usually indefinite subjects. Results on a preliminary evaluation are presented and discussed.

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

1. Introduction

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 (*pronouning*; [1]) or to zero (NP deletion) in order to avoid redundancy from repetition [2].

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

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

In sentence (1), the word *John* cannot occur in the second clause to avoid that the same entity be mentioned twice within the same sentence. This is made through prominalization. It is the sentence structure that determines, within limits, when the second mention of the entity will be named again or it will be referred to by a pronoun [3] or by zero.

In sentence (2) the words *he went* (in the second clause) may or not occur. The speaker chooses 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 (3), but not the opposite (4).

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

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

According to Harris [2], the phenomenon described above is the result of the reduction of a sentence in which the meaning of the source sentence stays unaltered.

Chomsky [4] proposed a typology of languages based of the obligatory expression of pronouns of its (facultative) reduction. This kind of languages is called by him as *pro-drop* (pronoun dropping) languages.

Halliday and Hasan [3] designate this kind of cohesion mechanism as *ellipsis*. According to them ellipsis is the omission of an item.

Finally, Mitkov [5] name this phenomenon of the omission of a word as *zero anaphora* or ellipsis. Accordingly, zero anaphors are 'invisible' anaphors, i.e. anaphors that do not appear to be in the sentence because they are not overtly represented by a word or phrase. On this study we adopted the same terminology used by Mitkov [5].

We focus on Brazilian Portuguese (BP) language. 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 vs. European Portuguese (EP):

(5) **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'

(6) *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'

(7) *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 (5) the NP *O João* 'John' cannot occur in the second clause because the same entity was already referred in the first clause. In sentence (6) 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 (7) 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 (7) 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 (8) 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 (9), the symbol Ø appears before the NP *o óvulo* 'the ovum'.

(8) *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"

(9) *Caso* Ø$_i$ *não seja fecundado, o óvulo*$_i$ *morrerá*
'If [the ovum] is not fertilized, the ovum will die'

1.1 Goal

The main goal of this paper 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.

Using this rule-based approach, we expect to improve the general performance of the Portuguese grammar [6] developed for Xerox Incremental Parsing (XIP) (under the collaboration between L2F[1] laboratory at INESC_ID[2] Lisbon and XRCE[3]) [7], namely by producing better dependency structures with reconstructed zeroed NPs for the 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.

2 Scope and Methods

2.1 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 or scope will also have to be made, and we will present in the appropriate place.

2.2 Sentences types

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

2.2.1 Coordinate sentences

A sentence is classified as coordinate when the clause does not carry any syntactic function in relation to another clause. Beside the clauses can hardly be the constituent's mobility, i.e. the sentence order cannot be changed [8].

(10) *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] Spoken Language Laboratory: https://www.l2f.inesc-id.pt/wiki/index.php/Main_Page
[2] Institute for System and Computer Engineering Research and Development in Lisbon: http://www.inesc-id.pt/
[3] Xerox Research Centre Europe: http://www.xrce.xerox.com/

(11) **Mas não foram de férias, o João e a Maria viajaram para o Sul*
*‘But they were not in vacation, John and Mary travelled to the South’

The main element that constitutes coordinate sentences is the conjunctions (named coordinating conjunctions) and the principal function of these conjunctions is explicit the link between the coordinated terms [8].

The coordinative sentences have three subtypes: additive, adversative and alternative. 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.

2.2.2 Subordinate sentence

The subordinate clauses considered on this study were: when the subordinate clause has a syntactic function in relation to the main clause (nominal subordinate clause) (12); or it expresses circumstantial events that modify the main clause (adverbial subordinate clause) (13)[4].

(12) *O João disse que não estava se sentindo bem*
‘John said that [he] was not feeling good’

(13) *O tempo mudou quando anoiteceu*
‘The weather has changed when it got dark’

2.2.3 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).

The nominal subordinate clause in which the NP subject can be zeroed is divided into three[5] types. 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

(14) *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

(15) *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 (dative cases) of the main clause

(16) *Fleury*$_i$ *insiste em que* Ø$_i$ *apenas deu despachos interlocutórios*
“Fleury insists that [he] only gave interlocutory orders”

(17) *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.

2.2.4 Adverbial subordinate clause

Adverbial subordinate clauses are characterized by exercising the syntactic function of adverb regarding 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 optional and the verb is in infinitive or in gerund or in participle mode.

The adverbial subordinate clause in which the NP subject can be zeroed is divided into six types[6]. This division is made based on the semantic information of the adverbial subordinate clause. The types are:

---

[4] The adjective subordinate clauses (relative clauses) 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 to discriminate in which cases the relative pronoun is the subject.
[5] Grammars consider four types, however the appositive subordinate clauses were not considered because the cases in which these phenomena occur are not many being unrepresentative.
[6] Some grammars consider nine types, but, in this dissertation, the comparative adverbial subordinate clauses, the conformative adverbial subordinate clauses and the proportional adverbial subordinate clause were not considered because the cases in which the NP subject can be deleted are not many being unrepresentative.

a) Conditional

(18) *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"

b) Causal

(19) *Como ela*$_i$ *a conhece bem,* Ø$_i$ *não fez nada*
"As she know her well, [she] did nothing"

c) Finality

(20) *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"

d) Concessive

(21) *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'

e) Time

(22) *Lula*$_i$ *evitava os debates quando* Ø$_i$ *liderava as pesquisas*
"Lula avoided the debate when [he] led the poll"

f) Consecutive

(23) *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"

2.2.5 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 [9]. For example:

(24) *O Pedro*$_i$ *queria* Ø$_i$ *ir ao cinema*
'Peter wanted to go to the movies'

(25) *O Pedro mandou lavar a louça*
'Peter asked to wash the dishes'

In the sentence (24), the subject in the subordinate clause is obligatorily coreferent to the subject of the main verb while in sentence (25) they cannot be coreferent. This information must be encoded in the lexicon so that it may be used in anaphora resolution. In section 2.5.6 we present a solution that integrates subcategorization information in zero anaphora resolution rules to deal with these cases.

2.3 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.

(26) *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 (26) 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 written in the first sentence.

2.4 Corpus

To our knowledge, there is no available corpus marked up with deleted subject NPs for Portuguese. 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 [10].

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.

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 [10]. 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 current content. In this table, there are the different genres texts – special report, news, chronicle, short stories and romance discriminated in the Text Types column; the number of the words that compound each genre – described 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 split

	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 (romance)	3,763	17%	2,277	18%	6.040	17%
Total	22,385		12,827		35.212	

The corpus was manually annotated[7]. The evaluation corpus was annotated separately and was only used for testing.

The Sentence corpus

The Sentence corpus consists on a set of sentences retrieved from the CETENFolha[8] corpus [11]. In addition, another set of sentences was specially constructed in order to test the rule of control verbs and the rule of attributes. This corpus was annotated following the same annotation guidelines used on the ZAC corpus.

2.5 Linguistically motivated rules

After a systematic linguistic analysis of the zero anaphora cases presented above, some general rules were defined and implemented[9].

2.5.1 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.

The rule that deals with coordinate sentences (Figure 1) 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.

Figure 1: 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)
```

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.

This is done by the following two rules:

---

[7] Too see the annotation guidelines, please refer to [10].
[8] Available at: http://www.linguateca.pt/
[9] 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.

Figure 2: Rule for coordinate NPs

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

     SUBJ[anaph0=+,pre=+](#2,#4)
```

Figure 3: 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 deeds to be duplicated.

2.5.2 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 4) 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 the second rule, there is a subordinate clause, indicated by the chunk `SC` (subclause). This is construed *grosso modo* by linking a subordinate conjunction to the fist finite verb to its right.

Figure 4: 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.

2.5.3 Anteposition of the subordinate clause

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

(27) *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 (27), 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 5).

Figure 5: 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 ',' [10] ; then a zero-anaphoric subject dependency is created and the subject is reconstituted from the subject of the first clause.

2.5.4 Infinitive adverbial subordinate clause

One of the most common cases of zeroed subject anaphor happens in infinitive[11] adverbial subordinate clauses. To solve these cases, the following rule has been developed:

Figure 6: 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 2.2.5 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.

2.5.5 Gerundive subordinate clause

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

(28) *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 7: Rule for the gerundive subordinate clause

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

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

(29) *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.

2.5.6 Control verbs and subordinate nominal clauses

As it was mentioned in section 2.2.5, 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:

(30) 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.

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 coreferencial constraints shown above. However the following rule (Figure 8) has been added for verbs with `s_pp_qufconj`[12] like ordernar 'to order':

---

[10] The requirement of comma was meant to limit the scope of the rule.

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

(31) *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 8: 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)
```

2.5.7 Attributes

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

(32) *O Pedro estava alegre*
'Peter was happy'

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

(33) *Ela um dia se casará e será muito infeliz*
'She will get married one day and will be very unhappy'

(34) *Branca de Neve*$_i$ *é tonta e* Ø$_i$ *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 silly for not have 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:

Figure 9: Rule for the attribute

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

     ATTRIB[anaph0=+](#3,#2).
```

Figure 10: 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 first rule simply extends the anaphoric argument subject of the `PREDSUBJ` dependency to the `ATTRIB` dependency (Figure 9). The second rule is slightly more complex for it checks on the other dependencies of the sentence without the copula verb (Figure 10).

3 Evaluation

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[13]. 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[14] and they are presented in Table 2.

---

[12] 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: *O Pedro pediu à Ana*$_i$ *que* Ø$_i$ *lavasse a louça* / 'Peter asked to Ana that [she] wash the dishes'

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

[14] 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.

Table 2: 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.

The most common errors fall mainly on three types sometimes connected: (a) POS tagging, (b) chunking and (c) dependency extraction, including ZA rules.

4 Conclusion

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 revolve zero subject anaphora in a fully integrated NLP chain [6].

To this end, a specific corpus, the ZAC corpus [10] 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, particular 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.

References

[1] Harris, Z. 1981. *Papers on Syntax*. Henry Hiz (Ed.). Dordrecht: D.Reidel Publishing Company.

[2] Harris, Z. 1991. *A Theory of Language and Information: A mathematical approach*. Oxford: Clarendon Press.

[3] Halliday, M.; Hasan, R. 1976. *Cohesion in English*. London: Longman.

[4] Chomsky, N. 1981. *Lectures on government and binding.* Berlin; New York: Mouton de Gryter.

[5] Mitkov, R. 2002. *Anaphora Resolution*. London: Longman.

[6] 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.

[7] 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.

[8] Matos, G. 2003. Estruturas de coordenação. In: Mateus, M.; Brito, A.; Duarte, I.; Faria, I.; Frota, S.; Matos, G.; Oliveira, F.; Vigário, M.; Villalva, A. *Gramática da Língua Portuguesa.* Lisboa: Caminho. pp: 551-592.

[9] 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

[10] 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.

[11] 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.