ZAC.PB: An Annotated Corpus for Zero Anaphora Resolution in Portuguese

Simone Pereira
University of the Algarve
Campus de Gambelas
P-8005-139 Faro, Portugal
simonecp@gmail.com

Abstract

This paper describes the methodology adopted in the construction of an annotated corpus for the study of zero anaphora in Portuguese, the ZAC corpus. To our knowledge, no such corpus exists at this time for the Portuguese language. The purpose of this linguistic resource is to promote the use of automatic discovery of linguistic parameters for anaphora resolution systems. Because of the complexity of the linguistic phenomena involved, a detailed description of the different situations is provided. This paper will only focus on the annotation of subject zero anaphors. The main issues regarding zero anaphora in Portuguese are: indefinite subjects, either without verbal agreement marks or with first person plural or third person plural verbal agreement; position of the anaphor relative to its antecedent, i.e. anaphoric and cataphoric relations; coreference chains inside the same sentence and spanning several sentences; and determining the head of the antecedent noun phrase for a given anaphor. Finally, preliminary observations taken from the ZAC corpus are presented.

Keywords

Anaphora resolution, zero anaphora, corpus linguistics, corpus annotation, syntax, Brazilian Portuguese.

1. Introduction

In many linguistic situations, redundant 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 [1].

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

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

Portuguese has a very rich verbal inflection, and the subject can easily be recovered through verbal inflection.

The grammatical rules governing NP deletion may vary among languages, even among different varieties of the 'same' language, as in the case of Brazilian (^{bp}) vs. European Portuguese (^{ep}). For example, the Portuguese equivalent for the examples (1.1)-(1.2) should be:

(1.3) **O João*$_i$ *foi à escola e depois o João*$_i$ *foi ao* ^{ep}*centro comercial/*^{ep,bp}*shopping*

(1.4) *O João*$_i$ *foi à escola e depois* (ε + ^{*ep,bp}*ele*$_i$) foi ao ^{ep}*centro comercial/*^{ep,bp}*shopping*

(1.5) *O João*$_i$ *foi à escola e depois* ao ^{ep}*centro comercial/*^{ep,bp}*shopping*

In the previous examples, the reduction of the verb imposes the subject NP deletion; otherwise it can be reduced, in Brazilian Portuguese, both to pronoun and to zero, while in European Portuguese only zero-reduction is allowed.

In order to correctly resolve zero anaphora [2], NLP systems require (a) the correct identification of the zero anaphor and (b) the correct identification of the antecedent of the zero anaphor[1]. Several strategies can be used to achieve this goal. For machine learning techniques, an annotated corpus is required.

This paper describes the methodology adopted in the construction of an annotated corpus for the study of zero anaphora in Portuguese, the ZAC corpus. To our knowledge, no such corpus exists at this time for the Portuguese language. The purpose of this linguistic resource is to promote the use of automatic discovery of linguistic parameters for anaphora resolution systems[2]. Our ultimate goal is to implement a module for zero anaphora resolution in the Portuguese grammar [3] developed under **X**erox **I**ncremental **P**arser (XIP) [4].

Because of the complexity of the linguistic phenomena involved, a detailed description of the different situations is provided. This paper will only focus on the annotation of subject zero anaphors. The main issues regarding zero anaphora in Portuguese are: indefinite subjects, either without verbal agreement marks or with first person plural or third person plural verbal agreement; position of the

[1] For clarity, *anaphor* is used to designate the pronoun, in NP reduction, or the syntactic slot left empty by NP deletion, while *anaphora* is a general term for the referential relation between the anaphor and its antecedent. It includes both *anaphora* proper, if the antecedent appears in a previous moment in discourse and *cataphora* if it appears after later moment.

[2] A similar corpus has been presented for Spanish [5] but in a different theoretical framework. A corpus for anaphora resolution has been produced for Brazilian Portuguese [6], but as far as we know only coreference chains between anaphors have been annotated, and no information was available for zero anaphors. Adaptation of the Mitkov algorithm [2] to Brazilian Portuguese pronoun resolution is given in [9].

anaphor relative to its antecedent, i.e. anaphoric and cataphoric relations; coreference chains inside the same sentence and spanning several sentences; and determining the head of the antecedent noun phrase for a given anaphor. Finally, preliminary observations taken from the ZAC corpus are presented.

2. Building the corpus

To our knowledge, there is no available corpus marked up with deleted subject NPs. Because of this lack on linguistic resources, an annotated corpus has been built for this study. The corpus consists on a set of full and partial texts retrieved from the web, and digitalized from books, encompassing several genres, namely journalistic and literary text from contemporary authors. The corpus is provided in text format, but the annotation adopted can be easily converted into other formats. Table 1 shows the breakdown per genre type of the ZAC corpus current content.

Table 1. Content of the ZAC corpus

	ZAC corpus	
Text types	words	%
Special Report	15.791	45%
News	1.769	5%
Chronicle	8.385	24%
Fiction (short story)	3.227	9%
Fiction (romance)	6.040	17%
Total	35.212	

3. Annotating the corpus

The corpus was annotated by two annotators working together. 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 and cataphora relations, respectively, and a word indicating the head of the antecedent noun phrase (NP).

3.1 Annotated cases

a) deleted subject

Only deleted subject of non-auxiliary verbs are to be marked. Verbal chains with auxiliary verbs whose subject has been zeroed count as a single verb form, hence there will be only one anaphor marked (3.1):

(3.1) *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 [males] having been exposed to insecticide

In coordinated clauses only the zeroed subject of explicit verb forms is marked (3.2):

(3.2) *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 (this situation is relatively rare in the corpus) (3.3):

(3.3) *"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 '`?`'.

On coordinated relative clauses, where the second relative pronoun has been zeroed, it is marked with the special notation `[0(que)=<X]`, where `X` represents the antecedent of the relative pronoun (this situation is not frequent in the corpus) (3.4):

(3.4) *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 processes can also occur by the modification of histones, the lines that involve the DNA and form a ball

b) noun phrases

For NPs whose head is a nominal determiner, for example *conjunto* 'set' (3.5) and *maioria* 'majority' (3.6), 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

(3.5) *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

(3.6) *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

In the case of compound nouns, only the head noun is to be referred to in the zeroed anaphor. Because of tokenization criteria we use, prefixed nouns are considered compound words (e.g. *ex-colegas* 'ex-partners') (3.7):

(3.7) *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, is to referred to by the form *gente* (3.8):

(3.8) *— 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', which will be referred to by the head noun *mundo* (3.9):

(3.9) *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 (3.10), syntactically non-analyzable, and half-frozen expression with infinitives (3.11) are left without notation:

(3.10) *[...] 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

(3.11) No decorrer das décadas*, no entanto, a população acabou se aprofundando na miséria.*

Over the decades, however, people just went deeper into poverty

Compound proper names (named entities, in majuscules) are considered a single token and therefore, will be referred to in the notation of zero anaphors. In the case of titles in apposition with proper names, the two elements are considered together as the head of that NP (e.g. Dona Marta 'Mrs Marta').

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.

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 (3.12):

(3.12) *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

c) indefinite subject

The indefinite subject is annotated as `[0=indef]` (3.13):

(3.13) `[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

Indefinite elliptical subject where there is a systematic ambiguity with first person plural *nós* 'we', will be specially noted `[0=1p]` (3.14):

(3.14) *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 resolved at this stage.

Sentences with zeroed subject and with the verb in the third-person plural will be annotated `[0=3p]`; this type of subject is systematically ambiguous between: a) indefinite subject with the particular (empowered) connotation; and b) a simple third-person plural, only context can disambiguate it (3.15):

(3.15) *"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, speaking 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=?]` (3.16):

(3.16) *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

d) impersonal subject

The impersonal subject is annotated as `[0=impers]`. This notation may cover different syntactic and semantic structures, such as: impersonal constructions[3] with verbs *ter* (in BP) and *haver* (both in BP and EP); meteorological constructions; temporal expressions [7].

3.2 Correference chains

A correference chain is established between a sequence of anaphors and their antecedent noun phrase. When the antecedent of a zero anaphor is in a previous sentence[4], the notation `[0=<<X]` is used even if the first element is in a fronted subordinate clause. The zero anaphor will be marked `[0=<<X]`, no matter how many sentences away it may be (3.17):

(3.17) *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

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) `[0=1p]` instances, the signs for anaphoric (<)/cataphoric (>) relation are not used.

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 later 0_1 is repeated.

In certain cases, a coreference chain can be determined among indefinite subjects; in this (relatively rare) situation, the coreference relation is marked `[0=<indef]`. The same happens with other indefinite subjects, such as the first-person plural (1p), and the third-person plural (3p).

3.3 Excluded cases

In the annotation, some cases were excluded.

a) adjectives

The subject of adjectives is only marked if they appear with their copula verb (e.g. *ser*, *estar*, 'to be') (3.18). Therefore the zeroed subjects of adjectives in apposition are not marked (3.19):

(3.18) *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

(3.19) *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

b) past participle

The past participle is considered as an ordinary adjective and its zeroed subject should be marked accordingly depending on the presence (3.20) or absence (3.21) of the copula verb.

(3.20) *Certamente* `[0=<marido]` estava *armado*

Certainly the husband was armed

(3.21) *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

[3] Impersonal constructions may also appear with a NP and a gerund (BP/EP) or a prepositional infinitive (only in EP): [0=impers] ^{bp,*ep}Tem/^{bp,ep}Há gente ^{bp,ep}fazendo/^{*bp,ep}a fazer isso 'There is people doing this'.

[4] The separators ';' and ':' are considered sentence boundaries, along with other common sentence separators ('.', '?', '!', etc.).

The past participle is considered a verbal form when it makes part of a compound tense with auxiliary verbs *ter* 'to have' (3.22) or (rarely) *haver* 'to there be' (3.23):

(3.22) *"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"

(3.23) *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

c) 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 (3.24). Otherwise gerundive adverbial clauses need the marking of zeroed subjects (cfr. (3.5)):

(3.24) *Luiz percebeu faíscas* saindo *de um poste à frente da casa*

Luiz saw sparks coming out of a pole in front of the house

d) 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 is not signaled (3.25):

(3.25) *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 (3.26):

(3.26) *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

e) direct speech, imperative, interrogative and exclamative sentences

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.

In much the same way, the zeroed subject of imperative sentences; direct, total (yes/no) or partial (*wh-*); interrogative sentences; question tags; and exclamative sentences, where the speaker or the addressee are integrated in the discourse, are not to be marked. 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 a zero anaphor (3.27):

(3.27) *— É essa casa aqui.* `[0=?]` *Estão ouvindo?*

— This here is the house. Are [you_3PL] listening?

f) causative operator verbs

On constructions of causative operator verbs [8] with restructured subject, the structurally zeroed slot of the subject of the dependent clause is not marked (3.28):

(3.28) *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

g) reduced, infinitive prepositional clauses

Reduced, infinitive prepositions clauses, usually resulting from the reduction of relative are treated as other relatives, that is, no zero anaphor is considered (3.29):

(3.29) *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).

4. Preliminary results

In this section we present preliminary results from the annotation process of the ZAC corpus.

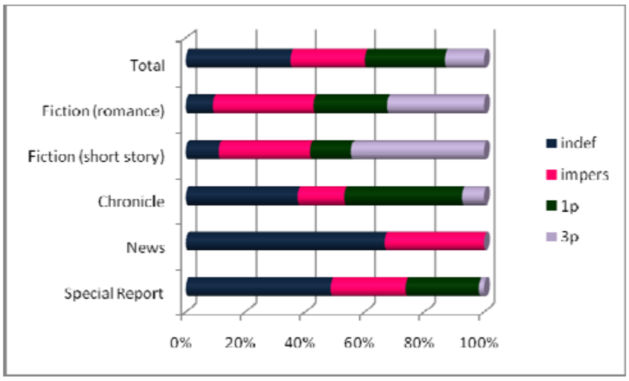
Table 2. Indefinite/impersonal subjects per genre

ZAC corpus						
Text types	words	total marks	indef	impers	1p	3p
Special Report	15791	538	81	42	41	3
News	1769	52	8	4	0	0
Chronicle	8385	395	41	17	43	8

Fiction (short story)	3227	146	4	11	5	16
Fiction (romance)	6040	358	7	26	19	25
Total	3512	1489	141	100	108	52

Table 2 presents the breakdown per genre of indefinite and impersonal subjects in the corpus. This type of subjects does not correspond to zero anaphors and their identification constitutes a linguistic challenge for any anaphora resolution system. Overall, they represent 401 (26.93%) from all zero subjects in the ZAC corpus. Figure 1 provides a comparative overview of this subject types.

Figure 1. Indefinite/impersonal subjects per genre



The 1p and 3p indefinite zero-subject types may be targeted by using the verbal inflection as a clue and in the absence of any other candidate antecedent NP; they still represent around 10% of the corpus zeroed subjects. Indefinite zeroed subjects, without 1p or 3p inflection associated, are harder to identify. Usually with verbs in the infinitive, they represent another 10% of the zeroed subjects. Finally, the identification of impersonal constructions (around 7% cases) heavily relies on the resolution of other syntactic issues such as auxiliary constructions and temporal expressions[5].

Table 3 presents the breakdown of anaphoric and cataphoric zero anaphora per genre and also distinguishes the anaphora with intra- (< , >) and intersentencial (<< , >>) antecedent.

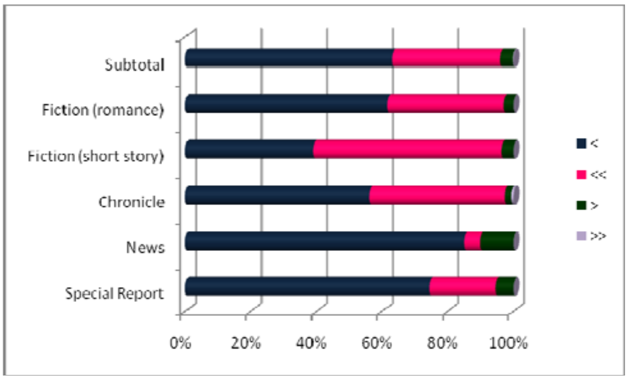
Table 3. Anaphora/cataphora breakdown per genre

ZAC corpus				
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 (romance)	171	99	8	0
Subtotal	680	355	41	2
Total	1035		43	

As one can see, cataphora is a relatively rare phenomenon, affecting a little over 3% of all anaphors in the corpus. Intrasentencial anaphora (<) represents 65% of all anaphors while intersentencial anaphora (<<) constitutes 34%.

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 2. News and special reports genres show clear predominance of intrasentencial anaphora (around 80 and 70%, respectively); fiction (romance) 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 2. Anaphora/cataphora breakdown per genre



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 a positive identification of antecedent NP is impossible represents 2.6%.

5. Future work

Based on these preliminary results we intend to develop a rule-based grammar for the identification of impersonal

[5] Since [2] do not provide explicit figures on this zero subject types it is not possible to compare our results with theirs. Nevertheless, it would be interesting to compare equivalent linguistics phenomena in both languages, Portuguese and Spanish.

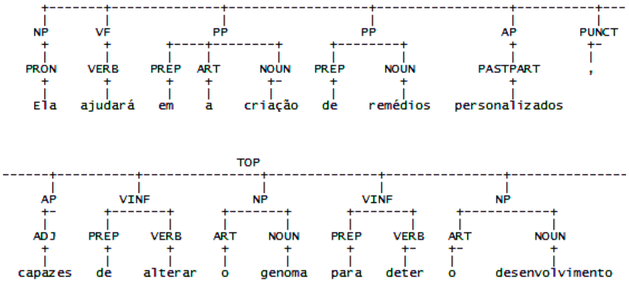
subjects that is to be integrated in the Portuguese grammar for XIP [3][4]. The temporal expressions have already been identified [7]. Auxiliary verbs involving verbs *ter* 'to have' and *haver* 'there be' are currently being implemented in XIP. It is likely that the reaming constructions of these verbs without explicit subject may be captured by rules of the XIP expressive formalism.

Secondly, we envisage a rule-base approach for the detection of the main syntactic configurations involving zero anaphors namely subordinate clauses.

Consider for example, sentence (3.19), renumbered below:

(5.1) *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*

Figure 3. Parse tree for sentence (5.1)



This sentence contains two prepositional phrases with infinitives (*de alterar* 'of changing' and *para deter* 'for stopping'). These phrases constitute two VINF chunks (Figure 3). Since there is no NP marked with a SUBJ[ect] dependency on those verbs yet, a rule could produce with some confidence the zero anaphor.

Once the rule-based approach attains its limits, we intend to explore the machine learning techniques described by [2] and [9].

6. Acknowledgements

Research for this paper was funded by a European Union grant, under the Erasmus-Mundus Program in the International Master in Natural Language Processing and Human Language Technology. The research has been supervised by Jorge Baptista, University of the Algarve.

7. References

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

[2] R. Mitkov. Anaphora resolution. UK:Longman, 2002.

[3] N. Mamede, J. Baptista, P. Vaz, C. Hagège. Nomenclature of chunks and dependencies in Portuguese XIP grammar (v. 2.1.). Lisboa: L2F-INESD-ID Lisboa (Internal Report), 2007.

[4] S. Ait-Mokhtar, J. Chanod, C. Roux. Robustness beyond shallowness: incremental dependency parsing. Natural Language Engineering 8 (2/3), pp. 121-144, 2002.

[5] L. Rello and I. Ilisei. A Comparative Study of Spanish Zero Pronoun Distribuition. Besançon: International Symposium on Data and Sense Mining, Machine Tanslation and Controlled Languages, pp. 209-214, 2009.

[6] S. Collovini, T. Carbonel, J. Fuchs, J. Coelho, L. Rino, R. Vieira. 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, 2007.

[7] C. Hagège, J. Baptista, N. Mamede. Portuguese Temporal Expressions Recognition: from TE characterization to an effective TER module implementation. 7th Brazilian Symposium in Information and Human Language Technology, SBC, 2009.

[8] M. Gross. Les bases empiriques de la notion de prédicat sémantique. Langages, 63, pp. 7-52. 1981.

[9] A. Chaves, L. Rino. The Mitkov Algorithm for Anaphora Resolution in Portuguese. A. Teixeira et al. (Eds.): PROPOR 2008, LNAI 5190, Springer-Verlag Berlin Heidelberg, pp. 51–60, 2008.