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

Simone Pereira
University of the Algarve, Portugal
simonecp@gmail.com

## Abstract

This presentation 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.

## 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) **John went to school and then John went to the mall*

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

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) - (2) should be:

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

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

*(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.

The purpose of this linguistic resource is to promote the use of automatic discovery of linguistic parameters for anaphora resolution systems. A similar corpus has been presented for Spanish [2] but in a different theoretical framework. A corpus for anaphora resolution has been produced for Brazilian Portuguese [3], 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 [4] to Brazilian Portuguese pronoun resolution is given in [5].

Our ultimate goal is to implement a module for zero anaphora resolution in the Portuguese grammar [6] developed under Xerox Incremental Parser (XIP) [7].

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.

## Building the corpus

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 | |

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.

## Annotating the corpus

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).

**1) Annotated cases**

a) deleted subject

✓ Only deleted subjects of explicit verbs are to be marked:

(3) *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:

(4) *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…

b) noun phrases

✓ In the case of compound nouns, only the head noun is to be referred to in the zeroed anaphor:

(5) *Para* `[0=>Ministério]` *tentar incentivar a criação de mais mestrados profissionais no País, o Ministério da Educação publica hoje uma 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

✓ Compound (frozen) expressions, syntactically non-analyzable (6), and half-frozen expression with infinitives (7) are left without notation:

(6) *[…] 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

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

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

✓ 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:

(8) *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]:

(9) `[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

d) impersonal subject

✓ The impersonal subject is annotated as [0=impers]:

**2) Coreference chains**

✓ When the antecedent of a zero anaphor is in a previous sentence, the notation [0=<<X] is used. The zero anaphor will be marked [0=<<X], no matter how many sentences away it may be:

(10) *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

Detailed annotation guidelines are provided in the full paper.

## Preliminary results

✓ Indefinite and impersonal subjects represent 401 (26.93%) from all zero subjects in 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 | 35212 | 1489 | 141 | 100 | 108 | 52 |

✓ cataphora is a relatively rare phenomenon, affecting a little over 3% of all anaphors in the corpus.

✓ intrasentencial anaphora (<) represents 65% of all anaphors

✓ intersentencial anaphora (<<) constitutes 34%.

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 | |

Figure 1. Indefinite/impersonal subjects per genre

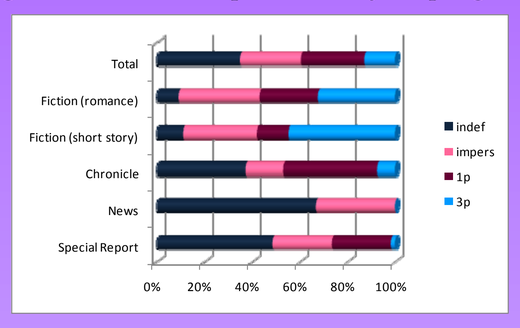
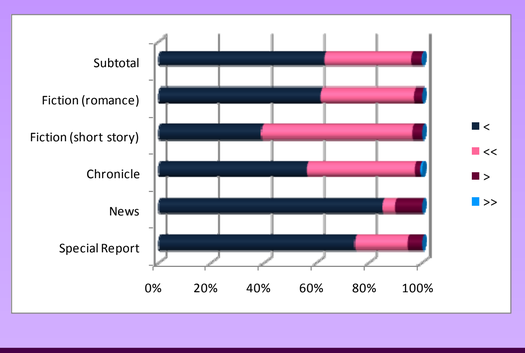



Figure 2. Anaphora/cataphora breakdown per genre




## Future work

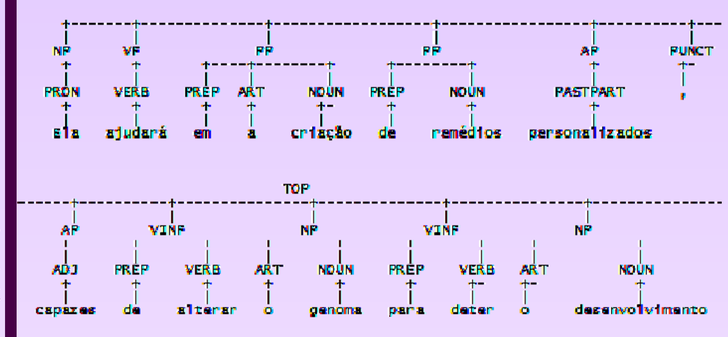
We envisage a rule-base approach for the detection of the main syntactic configurations involving zero anaphors namely subordinate clauses.

Consider for example, the sentence below:

*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

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.




## 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.

### References

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

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