ZAC: Zero Anaphora Corpus
A Corpus for Zero Anaphora Resolution in Portuguese

Jorge Baptista[1,3], Simone Pereira[1,3], and Nuno Mamede[2,3]

[1] Universidade do Algarve, Faculdade de Ciências Humanas e Sociais
Campus de Gambelas, 8005-139 Faro, Portugal
jbaptis@ualg.pt
[2] Universidade de Lisboa, Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[3] INESC-ID Lisboa/L2F – Spoken Language Lab
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
Nuno.Mamede@tecnico.ulisboa.pt

Abstract This paper describes a *corpus* of Brazilian Portuguese texts built in view of the construction of an Anaphora Resolution system, which is part of a fully-fledged Natural Language Processing system (STRING). The ZAC corpus is aimed at the resolution of the so-called *zero-anaphora*, that is, an anaphora relation where the anaphoric expression (or *anaphor*) has been zeroed The paper briefly discusses the linguistic issues in the process of zero anaphora resolution, and describes the annotation process in detail, as well as the main aspects of the anaphoric relations thus annotated.

Keywords Zero anaphora, Corpus, Brazilian Portuguese, Anaphora Resolution, Natural Language Processing

1 Introduction

The Natural Language Processing (NLP) task of Anaphora Resolution (AR) is critical for other NLP tasks, for example, for parsing and semantic role labelling, as well as for many applications such as machine translation, information extraction and question answering [1]. *Anaphora* is a major discursive device used to avoid repetition and increase the cohesion of the text, making the interpretation of a given sentence to depend upon the interpretation of previous elements [2, pp. 1701–12]. For example, in sentence (1), the (clitic) pronoun (the *anaphor*) *–la* refers to (is the *antecedent* of) the proper name *Amazônia* 'Amazonia' that appears in a previous moment in the discourse (thus, the relation between them being called *anaphora*). Besides contributing to the cohesion of the discourse, this anaphoric expression is also *co-referential*, *i.e.* both the name and the pronoun refer to the same extralinguistic entity, a geographic region in the real world:

(1) *Para salvar a **Amazônia** é preciso conhecê-**la**.* 'To save the *Amazonia* it necessary to know *her*'

Anaphora can also be classified according to the relative location of the antecedent and the anaphor: (a) *intrasentential anaphora*, if the antecedent is in the same sentence as the anaphor; (b) *intersentential anaphora*, if the anaphora relation is established across sentence boundaries; (c) *anaphora* proper, when the antecedent appears before the anaphor in the linear order of discourse; (d) *cataphora*, if that order is reversed. In addition to the immense amount knowledge that may be needed to perform anaphora resolution, the various forms that anaphora can assume make it a very challenging task, especially when one intends to "teach" computers how to solve anaphora. Machine-learning approaches to anaphora resolution, which constitute the main trend in current AR research, require large quantities of annotated corpora, where anaphoric relations are explicitly marked. Most of the previous work addressed pronominal anaphora, where the anaphor is a pronoun, as in (1). However, little work has been devoted to zero anaphora resolution and, to our knowledge, no corpus marked up with deleted subject noun phrases (NP) is available for Portuguese.

This paper presents the process of building a corpus with manually annotated zero-anaphoric relations in view of building an Zero Anaphora Resolution module [3] integrated in a fully-fledged Natural Language Processing system, STRING [4]. The paper describes the annotation process in detail, as well as the main linguistic aspects of the anaphoric relations thus annotated.

This paper is structured as follows: In the next section, we provide a brief overview of the major NLP approaches to anaphora resolution and current systems developed for Portuguese. Next, we present the corpus contents and then we describe in detail the main issues concerning zero anaphora, and the way they were annotated in the corpus. From this, the major results of the annotation process are presented. The paper concludes with some final remarks and perspectives for future work.

2 State of the Art

AR algorithms can be broadly classed into rule-based and machine learning approaches. Initially, it was the rule-based approaches such as Hobbs's algorithm [5] and Lappin and Leass's [6] resolution of anaphora procedure (RAP), which gained popularity. In the 1990s and 2000s, as people grew aware of the complexity of the job at hand, research started to be limited to specific types of anaphora in view of ultimately achieving better results. Dagan and Itai's collocation pattern-based approach [7]; Kennedy and Boguraev's parse-free approach [8]; Paraboni and Lima's research on Portuguese possessive pronominal anaphora [9]; Mitkov's algorithm [1] and Chaves and Rino's adaptation of Mitkov's algorithm for anaphora resolution in (Brazilian) Portuguese [10]; all these approaches brought new insights about AR and new ways to approach the task. Machine learning approaches to pronoun (and, in general, to anaphora and coreference) resolution [11, 12, 13, 14] have been an important direction of research. A corpus similar to ZAC has been presented for Spanish [15] but in a different theoretical framework.

In 2010, Pereira [3] presented a rule-based module for zero-anaphora resolution integrated in the STRING system [4]. The author reported a precision of 60.1%, a recall of 45.5% and a F-measure of 51.8%. In 2011, Nobre [16] implemented ARM 1.0, an adaptation of the Mitkov's algorithm for resolving Portuguese pronominal anaphora, achieving 33.5% F-measure, a value too low, compared to other state-of-the-art systems. Later, Marques [17] developed ARM 2.0 , an entirely new system, based on a hybrid, statistical and rule-based, approach, with a larger corpus, using a more complex annotation scheme [18]. The system reports 54.4% F-measure. It should be noted, however, that unlike previous work [1], no pre-processing of the corpus has been made, which renders the AR scenario more realistic. Both works [17, 16] only targeted pronominal anaphora, though.

3 Corpus Annotation

The Zero Anaphora Corpus (ZAC)[4] consists on a set of full and partial texts retrieved from the web, or digitalised from books, encompassing several genres and text types, namely journalistic and literary text from contemporary Brazilian Portuguese native-speaking authors, totalling 35,212 words. This corpus was split into two parts: the training corpus with 22,385 words (63.5%) and the evaluation corpus with 12,827 words (36.5%). Table 1 shows the breakdown per text type of the ZAC corpus current content.

Table 1. Breakdown of the contents of the ZAC corpus per text type.

	Training corpus		Evaluation Corpus		Full ZAC corpus	
Text type	**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 stories)	2,029	9	1,198	9	3,227	9
Fiction (novel)	3,763	17	2,277	19	6,040	17
Total	**22,385**		**12,827**		**35,212**	

The corpus was jointly annotated by two linguists, who revised and discussed each other's work, so that each annotation one of them encoded was always checked by the other annotator. Because of this methodology, no inter-annotator agreement measure can be provided. A set of very detailed annotation guidelines [19] were produced to help the annotation process and render it more consistent. For lack of space, only an outline of these guidelines is presented here.

The annotation of zero anaphora consisted, basically, in inserting a tag for the zero *anaphor* with the form '`[0=<x]`' in the empty slot of the zeroed constituent, linking it to its immediate *antecedent* (`x`) and determining whether it

[4] https://string.l2f.inesc-id.pt/w/index.php/Corpora [last access: 31-05-2016].

appeared *before* '<' (anaphora proper) or *after* '>' the anaphor (cataphora). Inter-sentential anaphora is marked with double arrows '<<' and '>>', irrespective of the number of intervening sentences.

Briefly, the following linguistic situations were encoded. In coordinated clauses, only the subject of explicit verbs under coordination are marked. Clausal antecendents are indicated by their main verb (5).

(5) *"**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'

On coordinated relative clauses, where the second subject relative pronoun has been zeroed, this should be marked but with the special notation `[0(que)=<X]`, where `X` represents the antecedent of the zeroed relative pronoun, as seen in (6):

(6) *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'

Zeroed subjects of gerundive adverbial subclauses are also marked (7):

(7) *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'

In the case of antecedent noun phrases with nominal determiners (*e.g.*, *milhão* 'million' (8) and the percentage expression *por cento* 'percent' (9) or its corresponding symbol '%'), it is the semantic head noun (syntactically, a complement of the determiner), that is chosen as the antecedent:

(8) *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'

(9) *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 of a NP has been zeroed in front of determiners, the determiner is then taken as the head noun of that NP and functions as antecedent for the following zero anaphor (10); in this way, the zero anaphor always refers to its syntactic antecedent (and not to the antecedent noun itself, which can be very far way from the current sentence). This approach is also adopted for nominal determiners like *maioria* 'majority' (11):

(10) *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'
(11) *Dos 25% restantes, a* **maioria** *pediu desculpas,* `[0=<maioria] explicando que [0=<maioria]` *tinha marcado de* `[0=<maioria]` *sair com a namorada.* 'From the remaining 25%, the majority appologized, explaining that [they] already had a date with their girlfriend'

The annotation of *zero indefinite subjects* is somewhat different, since they do not constitute anaphors, but may hinder significantly the anaphora resolution process. *Zero-indefinite* subjects are marked as `[0=indef]` (12):

(12) `[0=indef]` *Nascer com patrimônio genético idêntico não significa que as pessoas crescerão* `[0=<pessoas]` *tendo corpo, mente e doenças iguais.* 'To be born with identical genetic heritage does not mean that people will grow up having a similar body, mind and diseases'

First person plural indefinite subject, where there is a systematic ambiguity with zeroed pronoun *nós* 'we', is specially noted `[0=1p]`. In the example (13) , the first person plural may correspond to: (a) a real plural, referring to the speaker and his/her team of researchers; (b) the so-called 'modesty' plural, referring to the (singular) speaker; or (c) the indefinite (generic) subject, referring to the scientific community as a whole. Naturally, such ambiguities cannot be solved at this stage. Similarly, sentences with the indefinite third person plural zeroed subject, where the verb in the third person plural, is annotated `[0=3p]`, as in (14). This type of subject is systematically ambiguous between the indefinite subject and a simply zeroed third person plural pronoun *eles/elas* 'they', so that only context can disambiguate it:

(13) *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'
(14) *"Ainda* `[0=3p]` *estão fazendo isso lá embaixo",* `[0=<<Zé Lopes]` *acrescenta* (...) ' "[They] are still doing it down there," [Zé Lopes] adds'

The *impersonal subject* is annotated `[0=impers]`. This notation may cover different syntactic and semantic structures, such as *meteorological* constructions (15); and *impersonal* constructions with *haver* 'to there be' (both in Brazilian and European Portuguese)(16), or *ter* 'to have' (only in Brazilian Portuguese) (17):

(15) — *Nossa!* `[0=impers] Esfriou!` '— Wow. It got cold!'
(16) *"*`[0=impers]` *Há uma perigosa tendência a* `[0=indef]` *fazer correlações entre etnia, crime e predisposição genética"* ' "There is a dangerous tendency to establish correlations between ethnic origin, crime and genetic predisposition" '
(17) `[0=impers]` *Tem gente* `[0=<gente]` *fazendo isso.* 'There is people doing this'

Finally, the subject of adjectives (and past participles when used as adjectives) is only marked if they appear with their copula verb, therefore the zeroed subjects of adjectives in apposition, as *capazes* 'capable' in (18), are not marked:

> (18) *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 personalised medicine, capable of altering the genome in order to halt the development of diseases and mental disorders'

To conclude, some exceptions. *Topicalization* structures, *cleft sentences* with *ser ... que*, and other forms of focusing sentence elements involving changes in basic word-order are not marked and the syntactic position left empty by the moved constituent is not signaled. In the case of *direct speech* (for example, in interviews) the first person subject and the second person, if zeroed, are not marked. The zeroed subject of imperative sentences; direct, total (yes/no) or partial (*wh-*) interrogative sentences; question tags; and exclamative sentences are not to be marked, either. For lack of space we do not provide examples of such sentence types here.

4 Results

In this Section we present some of the main results from the annotation process. On the one hand, Table 2 presents the distribution of zero anaphors, zero-indefinite, impersonal constructions, 1p- and 3p-indefinite constructions. One can see that indefinites and impersonal constructions represent 26% of the corpus zero subjects, thus they constitute a serious hindrance to anaphora resolution. On the other hand, Table 3 shows the distribution of the anaphora/cataphora and intra-/inter-sentential distinctive types. Only 4% of the anaphoric relations correspond to instances of cataphora. The cases of intra-sentential anaphora represent 66.9% of the tags. It is noteworthy that in 53.8% cases of anaphora proper, the antecedent can not be found in the same sentence as the anaphor.

Table 2. Breakdown of zero-anaphors, impersonal and zero-indefinite subjects

	ZAC corpus					
Text Type	**zero**	**indef**	**impers**	**1p**	**3p**	**Total**
Special Report	371	81	42	41	3	538
News	40	8	4	0	0	52
Chronicle	286	41	17	43	8	395
Fiction (short stories)	110	4	11	5	16	146
Fiction (novel)	281	7	26	19	25	358
Total	1,088	141	100	108	52	1,489
Total (%)		0.73	0.09	0.07	0.07	0.03

Table 3. Distribution of the anaphora/cataphora and intra-/inter-sentential anaphora.

ZAC corpus				
Text types	<	<<	>	>>
Special Reports	275	74	20	0
News	34	2	4	0
Chronicle	156	115	5	2
Fiction (short stories)	44	65	4	0
Fiction(novel)	171	99	8	0
sub-total	680	355	41	2
sub-total (%)	0.631	0.329	0.038	0.002
Total	**1,035**		**43**	
Total (%)	**0.960**		**0.040**	

5 Conclusions and Future Work

This paper presented a corpus with manually annotated zero-anaphoric relations, as well as other related phenomena with direct bearing in the anaphora resolution process of zero anaphora, namely impersonal and zero-indefinite subject constructions. A set of annotation guidelines was produced [19], and briefly presented here, to better target the linguistic phenomena and provide consistency to the annotation process. To the best of our knowledge, this is the first corpus annotated for this type of phenomena for Portuguese. Results show that zero-indefinites constitute up to $\frac{1}{4}$ of the tags, which significantly complicates the AR process, while cataphora has only less than 5% frequency. Based on this corpus, a rule-based module for anaphora resolution has already been developed by Pereira [3, 20] and integrated in the Portuguese grammar of STRING system [4]. The evaluation of this module reported a 0.60 precision, 0.46 recall and 0.52 F-measure. Later, Marques [17] developed the ARM 2.0 hybrid AR module, currently used in STRING, but only targeting pronominal anaphora.

In the future, we expect to expand the ZAC corpus in order to include European Portuguese texts and to use machine learning techniques to improve the zero anaphora resolution in STRING. We also envisage to integrate pronominal and zero anaphora phenomena into a single, unified and coherent, AR module.

Acknowledgment

This work was partially supported by the European Union Erasmus Mundus Master Courses program (ref.EMMC 2008-0083) and by national funds through FCT–Fundação para a Ciência e a Tecnologia, ref. UID/CEC/50021/2013.

References

1. Mitkov, R.: Anaphora Resolution. Pearson (2002)
2. Mendes, A.: Organização textual e articulação de orações. In Paiva Raposo, E., Bacelar do Nascimento, M., Mota, A., Segura, M., Mendes, A., eds.: Gramática do Português. Volume 2. Fundação Calouste Gulbenkian, Lisboa (2013) 1691–1755

3. Pereira, S.: Linguistics Parameters for Zero Anaphora Resolution. Master's thesis, Univ. Algarve/Univ. Wolverhampton, Faro and Wolverhampton (2010)
4. Mamede, N., Baptista, J., Diniz, C., Cabarrão, V.: STRING: A Hybrid Statistical and Rule-based Natural Language Processing Chain for Portuguese. In: 10th Conference on Computational Processing of Portuguese. PROPOR '12 (Demo Session), Coimbra, Portugal (2012) `https://string.l2f.inesc-id.pt//`.
5. Hobbs, J.R.: Resolving Pronoun References. Lingua **44** (1978) 311–338
6. Lappin, S., Leass, H.J.: An Algorithm for Pronominal Anaphora Resolution. Computational Linguistics **20**(4) (1994) 535–561
7. Dagan, I., Itai, A.: A Statistical Filter for Resolving Pronoun References. In Feldman, Y.A., Bruckstein, A., eds.: Artificial Intelligence and Computer Vision. Elsevier Science Publishers B.V. (1991) 125–135
8. Kennedy, C., Boguraev, B.: Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. In: Proceedings of the 16th International Conference on Computational Linguistics. COLING '96, Copenhagen, Denmark, John Wiley and Sons, Ltd (1996) 113–118
9. Paraboni, I., Strube-de-Lima, V.L.: Possessive Pronominal Anaphor Resolution in Portuguese Written Texts. In: Proceedings of the 17th International Conference on Computational Linguistics. COLING '98, Montreal, Québec, Canada, Association for Computational Linguistics (1998) 1010–1014
10. Chaves, A.R., Rino, L.H.: The Mitkov Algorithm for Anaphora Resolution in Portuguese. In: Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language. PROPOR '08, Aveiro, Portugal, Springer-Verlag (2008) 51–60
11. McCarthy, J.F., Lehnert, W.G.: Using Decision Trees for Coreference Resolution. In: Proceedings of the 8th International Joint Conference on Artificial Intelligence. IJCAI '95, Montreal, Québec, Canada, Morgan Kaufmann Publishers Inc. (1995) 1050–1055
12. Cardie, C., Wagstaff, K.: Noun Phrase Coreference as Clustering. In: Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. EMNLP/VLC '99, College Park, Maryland, USA (1999) 82–89
13. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics **27**(4) (2001) 521–544
14. Rahman, A., Ng, V.: Supervised Models for Coreference Resolution. In: Proceedings of Empirical Methods in Natural Language Processing. EMNLP '09, Singapore, Association for Computational Linguistics (2009) 968–977
15. Rello, L., Ilisei, I.: A comparative study of Spanish zero pronoun distribuition. In: International Symposium on Data and Sense Mining, Machine Tanslation and Controlled Languages, Besançon, France (2009) 209–214
16. Nobre, N.: Resolução de Expressões Anafóricas. Master's thesis, Universidade Técnica de Lisboa - Instituto Superior Técnico (2011)
17. Marques, J.: Anaphora Resolution in Portuguese: A Hybid Approach. Master's thesis, Universidade de Lisboa - Instituto Superior Técnico (2013)
18. Marques, J., Baptista, J., Mamede, N.: Anaphora Annotation Guidelines. Technical report, L2F - Spoken Language Laboratory, INESC-ID Lisboa, Lisboa (2013)
19. Pereira, S., Baptista, J.: Zero anaphora corpus annotation guidelines. Technical report, L2F - Spoken Language Laboratory, INESC-ID Lisboa, Lisboa (2009)
20. Pereira, S.: ZAC.PB: An annotated corpus for zero anaphora resolution in Portuguese. In: Student Research Workshop in conjunction with RANLP-09, Borovets, Bulgaria (2009) 53–59