The annotation coreference task at IberEval'2017: the experience of CLUL/UE

Amália Mendes[1], Sandra Antunes[1], and Paulo Quaresma[2]

[1] Center for Linguistics of the University of Lisbon, Portugal
[2] Computer Science Department, University of Évora, Portugal
amaliamendes@letras.ulisboa.pt, sandra.antunes@gmail.com, pq@uevora.pt

Abstract. In this paper the process of coreference annotation in Portuguese texts in the context of a task of IberEval 2017 is described and the main observed problems are discussed. The work was done by a team of researchers from the Centre for Linguistics of the University of Lisbon (CLUL) and from the Computer Science Department of the University of Évora (UE). Due to time constraints and the complexity of the task, only researchers from CLUL were able to finish successfully the annotation process. The main problems are presented and discussed and some possible solutions are proposed. Nevertheless, the obtained results are similar with the overall results of the task.

1 Introduction

We report here our annotation experience in the scope of the coreference annotated corpus task at IberEval 2017. The first task was for each team to select a set of texts to be made available for annotation. The texts selected by the CLUL/UE team are taken from the LE-PAROLE Corpus, a 3 million words corpus of European Portuguese from different genres that was compiled for the LE-PAROLE project as the Portuguese counterpart of a set of comparable corpora of 20 European languages to be made available free of copyrights. For each language, a subset of 250.000 words was also annotated for POS and manually revised [5]. We selected texts from this corpus to ensure that the texts would be cleared for copyright issues and that the results of the coreference annotation task could be freely distributed to the community. Another reason was that this corpus allowed us to dispose of texts from a set of different genres. For the coreference annotation task, texts were meant to have a maximum of 1200 words. We selected the shortest texts of newspapers from the LE-PAROLE and, when the texts were longer, we adjusted the length of the document.

The annotation was performed using the editor CorrefVisual [6], developed by the Group of Natural Language Processing PLN-PUCRS at Pontifícia Universidade Católica do Rio Grande do Sul. We first annotated a text sent by the organizing committee for training and to report any problem that we might encounter with the editor. The result of the annotation of this text was then sent back to the organization of the task. This first stage of the task was very important to get used to the editor and to the guidelines, and for a first impression of the task.

The goal was then for each element of our team to annotate 10 texts, that were selected out of the set of texts sent by all the teams. Annotator1 (SA) annotated the full 10 texts, annotator2 (AM) annotated only 6 texts due to time constraints. All other annotators failed to finish the task: two of them due to difficulties with the computational process (they were not able to install and to run the annotator) and two of them due to time constraints and the linguistic complexity associated with the task (these annotators background was computer science).

Our experience annotating with the CorrefVisual editor is reported in section 2, some issues in identifying coreference relations are discussed in section 3, as well as our interannotator agreement in section 4, and some final remarks are presented in section 5.

2 Working with the CorrefVisual editor

The tool CorrefVisual [6] runs in Java and allows the edition of texts previously annotated with CORP [3, 2], a nominal coreference resolution tool for the Portuguese language. Two different operating systems (windows and ios) were used during the annotation and the tool worked well on both systems. It is important, however, to point out that two annotators were not able to install CorrefVisual in their computers due to problems with the operating system and Java versions. As they were not able to obtain technical help in time, they were not able to successfully finish the annotation.

We received a short description of the tool and its functionalities that was extremely useful to get acquainted with the editor and how to proceed with the annotation[3]. We also received guidelines with a description of the task, an explanation of the concept of coreference and examples of coreference chains and also negative examples. The guidelines were clear and well structured but, of course, considering the complexity of the task, we encountered several cases that were not considered in the guidelines and that we will discuss in section 3. Moreover, two annotators, having a good knowledge in computer science but a weak linguistic background, failed to understand all the concepts and implications of the task and they were not able to annotate the texts. This point clearly shows the high difficulty and the requirement of very specific skills for this task.

The two successful annotators in our team worked independently on their annotations and didn't discuss the work among them. They relied solely on the guidelines that were made available by the organizing team. This was done on purpose to properly evaluate interannotator agreement.

The main problem that was encountered with the editor was the delimitation of phrases: when changing the length of a phrase, the editor returned an error message saying that more than one phrase was selected, even when only one phrase was highlighted. After reporting on this issue, there was information to use the ESC key but making sure that at least one free phrase was selected,

[3] http://www.inf.pucrs.br/linatural/wordpress/index.php/recursos-e-ferramentas/correfvisual/

otherwise the unselection wouldn't work. This solved the problem and it was then easy to unselect all current selections (even invisible ones) and proceed with the selection and manipulation of a single phrase.

The NPs are not editable in the editor, and they have to be selected out of the list of NPs that have been automatically identified in the preprocessing stage. Some of the NPs were automatically attributed to a reference chain and the remaining ones were listed in a separate window. The annotator task was to verify the contents of each suggested reference chain and to modify it when necessary, using the list of NPs in the separate window (free nominal phrases).

The NPs in the reference chains were frequently inadequate and in this case it was necessary to remove the NPs to the list of free nominal phrases (by dragging it to the window) or to manipulate the NP to obtain the right length. When the NP that was aimed for was not available in the list of free nominal phrases, another NP close in the context had to be selected and its length manipulated until it fitted. In selecting a different NP, it was crucial to make sure that it would not be required for another reference chain.

For instance, there are two different reference chains in *Após a sua constituição*: the NP refers to one reference chain and the possessive pronoun to another one. In order to identify the possessive separately, one had to choose a free nominal phrase close by in the context and remove and add tokens until the possessive was correctly identified. In other cases, it was not possible to find a sequence, among the free nominal phrases available, that could correctly capture the NP that we were aiming for. For instance, according to the guidelines, the apposition should be identified as a separate referential expression, as in *o feiticeiro (o psiquiatra colectivo ou o Moreno de então)*, but there was no free nominal phrase that would single out the part between curved brackets.

To identify the correct NP in the list of free nominal phrases, the annotator could use the function Search nominal phrases, that would highlight the NPs containing the sequence that was queried. This was very helpful for the task. Several NPs could be highlighted as the result of the search (although not visible in the screen) and the annotators had to remind to check the results of the search before dragging an NP to the reference chain box, otherwise all the highlighted NPs would be dragged together. The identification of the correct NP in the list of free nominal phrases was anyway time consuming, especially when several NPs had very similar content. This involved checking them one by one in the context before selecting the right one.

Some of the problems that we experienced in the identification of the nominal phrases are due to the preprocessing of the texts, namely tokenization and the grouping of tokens as named entities. For instance, titles were grouped together with the first token of the following sentence as one named entity, as in *Hospital de Castelo_Branco_O_Hospital_Distrital* and also *Linha de o Corgo_Está*. Some named entities that were automatically identified contained more lexical material than required. For instance, in the sequence *Extinção_do_Gabinete_de_Planeamento e de Coordenação_do_Combate_à_Droga*, we couldn't eliminate the first two tokens *Extinção_do* and had to select the whole sequence. The tokenization pro-

cess wrongly treated the accusative pronoun *nos* as the contraction of *em* and *os*, in example (1). Consequently, we included both tokens as part of the reference chain.

(1) Para *em os* explicarmos temos de ir buscar a ponta a o princípio de o século (file dn81701)

The post-verbal accusative or dative clitic is usually treated as an independent token that can be identified separately as part of a reference chain (as in (2-a)), but in some cases, the tokenization didn't separate verb and clitic, and both tokens had to be selected as part of the chain, as illustrated in (2-b).

(2) a. o homem não devia obedecer a a natureza , mas sim vencê *–la* (dn88218).
b. desafiar e vencer a natureza *contrariando-a* (dn88218)

In several cases, the manipulation of the length of the NP would create an incorrect tokenization by including parts of the previous or following token. For instance, when modifying a sequence to obtain a single token *que*, it included the first letter of the following token *hoje*. The result of the annotation is *que h* and it couldn't be corrected.

3 Linguistic issues in annotating coreference

Near Identity

In many contexts, it is difficult to establish with absolute certainty that two NPs are coreferent [4]. For instance, in the initial training phase, we decided to treat as coreferent the NP *os primeiros cães domésticos* and the NP *cães domésticos* that occurs in the larger NP *fósseis de cães domésticos*. It can be debated whether the two are coreferent: although the second NP refers to fossils which are consequently old, it might not refer exactly to the fossils of the first domestic dogs. The two NPs were treated as coreferent to avoid dividing the data into many reference chains. This brings about the question of Near-Identity, which will be treated according to the level of granularity and the general goals of the annotation. Another example from the training phase is the NP *diversidade genética* and the noun *diferenças* that were treated as coreferent because the differences were interpreted contextually as genetic differences. This reference chain was already automatically pre-identified in the CorrefVisual editor.

In another case, we annotated two near corereferent NPs as part of different reference chains. In example (3), *ser humano* and *a humanidade* are treated as non coreferent due to the explicit mention of their different scope in the context, although they appear to be used in the rest of the text as synonyms.

(3) *o ser humano* e, por extensão, *a humanidade* (dn81201)

Nominal phrases may be lexically distinct but very similar in terms of their reference, as in examples (4-a) and (4-b), where *estações de recolha* and *estações*

meteorológicas refer to the same entity and *à escala mundial* and *o planeta* refer to the same scope of the network. We considered the two NPs as part of the same reference chain.

(4) a. a rede de estações de recolha a a escala mundial (pu92214)
b. a rede de estações meteorológicas de o planeta (pu92214)

The following case raises even more questions about what can be considered as coreferent. In example (5-a), the nouns *modelos matemáticos* and *computadores* are modified by an adjectival phrase with very specific lexical material. In example (5-b), the same nouns are modified by the less informative adjective *melhores*. Could we consider that the better models and computers that the second example mentions are the ones capable of modelizing the weather and the meteorological conditions? We consider that it is indeed the case, based on the context, and annotated as coreferent.

(5) a. Há alguns anos , faltavam estações de observação e não havia modelos matemáticos nem computadores capazes de modelizar o clima e as condições meteorológicas (pu92214)
b. Considera que os principais problemas consistem em a falta de dados de base , de melhores modelos matemáticos , de melhores computadores (pu92214)

Quantification

The quantification of the NPs raises many questions regarding the annotation of coreference. The NPs *descargas elétricas atmosféricas* and *um raio*, in (6-a) denote the same type of entity in the text and differ in terms of their register (more vs. less specialized). Although the first is in the plural form and the second in the singular, they both have a generic reading that points to a case of coreference (or near identity). Compare, however, (6-a) with (6-b): in the second sentence, the NP also has a generic reading but it is quantified. The question is whether it should be included in the same reference chain. However, quantified NPs are not considered coreferent: even if they denote the same type of entity, they refer to a specific subset of those entities. These issues should be made explicit in the annotation guidelines.

(6) a. Lago de Maracaibo , em a Venezuela , apresenta a concentração mais elevada de *descargas elétricas atmosféricas* de o mundo . Em algum lugar de o mundo está caindo *um raio* em este momento . (Texto11.txt)
b. 44 descargas elétricas atmosféricas a cada segundo (Texto11.txt)

Modality

The presence of epistemic modal markers (i.e, lexical markers that express values such as uncertainty, possibility) raises issues in terms of the annotation of coreference [1]. The same entity is referred to through the view of a different source

in examples (7-a) and (7-b) (the NPs are in italic, while the modal marker is underlined). This would mean that the coreference exists only for this specific source: for instance, *aquela organização do trabalho* and *um dos primeiros passos num caminho que tende a levar longe* have the same reference for the source, namely *o sacrossanto poder do patrão na empresa*, but it is clear that the author of the text disagrees with this view. Coreference would then be dependent on the view of each source.

(7) a. os depoimentos de quem viveu *a cadeia* , desde o velho Navel a o recente Haraszti , deixam -nos a impressão de *um trabalho destruidor...*" (Texto11.txt)
b. *Aquela organização de o trabalho* , a o conferir poderes a os trabalhadores sobre as condições de o trabalho , é vista por o sacrossanto poder de o patrão em a empresa como *um de os primeiros passos em um caminho que tende a levar longe* . de aí resistências (Texto11.txt)

In example (8), modality does not refer to the viewpoint of another source. Modality is expressed by a conditional clause and the equivalence between the two NPs is only valid if the condition applies. In this specific case, the condition is to be pragmatically understood as a goal, so that modality doesn't seem to affect the existence of coreference.

(8) *É o principal objectivo de a OMM* e é *a coisa mais sensata a fazer* se queremos compreender o fenómeno de o aquecimento global e de as suas implicações.

Embedded and coordinated NPs

One of the first issues faced during the annotation was whether an embedded NP could be part of another reference chain. For instance, the NP illustrated in (9) would refer to the two reference chains indicated in the example. We treated embedded NPs as part of other reference chains, when applicable.

(9) o estudo das sociedades primitivas (dn81201)
reference chain 1: o estudo das sociedades primitivas
reference chain 2: as sociedades primitivas

The same issue arose in what concerns coordinated NPs. For instance, could the NP *os primatas* in (10) be included in the reference chain *os antropóides*? We treated coordinated and embedded NPs similarly.

(10) Os animais inferiores , *os primatas* e o homem primitivo constituem uma linha evolutiva que se revela sempre mais complexa , até desembocar em o salto qualitativo que é a civilização . (dn81201)

Modified NPs

Another issue is whether a modified nominal head can be considered independently of its modifiers. For instance, in (5-a) the complex nominal head *modelos matemáticos* and the head *computadores* are modified by an adjectival phrase. We considered that the modifier had to be included since it restricts the reference of the nominal.

Implicit content

The antecedent of an anaphoric element can be implicit in the context. For example, the meetings refered in (11-b) are the meetings of the Comission refered in (11-a), so this entity is implicit in the NP: essas reuniões [da Comissão Intergovernamental de Negociação]. The question is whether the NP in (11-b) should be considered as part of the reference chain of the entity *Comissão Intergovernamental de Negociação*.

(11) a. Uma Comissão Intergovernamental de Negociação (pu92214)
b. Essas reuniões

An NP can refer to information that is scattered in the previous context, as the NP *todas essas estações de observação que é preciso montar ou reactivar*. The demonstrative relates to information showed in italic, but there is no clear NP that could be considered coreferent with the underlined NP and such cases were not annotated:

(12) *A falta de estações de recolha de dados* é a que nos parece de maior importância superar , para permitir melhorar a fiabilidade de as previsões . Quanto a os modelos que se usam actualmente , são muito rigorosos e não nos parecem responsáveis por a imprecisão de as previsões : é a qualidade de os dados que se fornecem a o computador que as condiciona . em uma previsão de 24 horas para um país de a Europa ocidental , por_exemplo , *é preciso ter dados de toda a Europa , de uma grande parte de África , etc.* Se quisermos uma previsão a quatro dias , já se torna *necessário cobrir cerca_de metade de o globo* . E , para uma previsão além_dos os quatro ou cinco dias , *são precisos dados de todo o globo* . Mesmo que se tenha uma densidade óptima de estações de observação em Portugal , não podemos esperar uma boa previsão , nem_sequer para Portugal , se *o resto de o mundo não estiver bem coberto* . Existem mais de dez mil estações terrestres de observação em o mundo e mais de 1500 estações atmosféricas . (...)
Como espera financiar todas essas estações de observação que é necessário montar ou reactivar?

Relatives

Restrictive relatives were included in the NP because they contribute to establish the reference of the NP. There could be two possible annotations of such cases,

illustrated in (13-a) and (13-b). Option 1 (a single NP) was preferred due to the fact that the restrictive relative is included in the NP and is crucial for the reference, but their was some hesitation in the annotation.

(13) a gorila a quem M. Patterson conseguiu...
  a. [a gorila a quem M. Patterson conseguiu...]
  b. [a gorila] a [quem]

Length of the NPs

Other constructions beside relative clauses raise questions about what to include in the NP of a reference chain. For instance, in the case of the context illustrated in (4-a), the issue was whether *à escala mundial* should be included in the NP or not. The fact that a similar NP, illustrated in (4-b), occurred in the text lead to the selection of the whole sequence. In contexts such as (14), the parenthetical segment wasn't included because it is not essential to the reference of the NP (just as a non restrictive relative would also be left out).

(14) a definição e concretização de uma estrutura associativa empresarial sólida e eficaz, *essencialmente de base regional* (dn81625)

4 Results and interannotator agreement

Based on two files annotated by both annotators (pu92214 and dn81201), the interannotator agreement reached a moderate kappa value of 0.40. We inspected our results and compared the annotation of these files. In both files, Annotator2 treated more reference chains than Annotator1: 65 vs. 40 chains for pu92214 and 75 vs. 43 chains for dn81201. There is an average of 37 common reference chains in the two annotations. The number of NPs in these common chains is very similar among the two annotators. The differences lie mostly in the annotation of embedded NPs, and we believe that this could be easily improved with explicit mention in the guidelines. There is also a difference in the number of pronominal elements in the chains (demonstrative, personal, possessive and relative pronouns) and in the length of some NPs. For example, in the file dn81201, 22 reference chains that were identified by Annotator2 (and not by Annotator1) involve relative pronouns, as can be observed in (15-a) and (15-b).

(15) a. homem revelou aquela alma espiritual que hoje parece ser suficiente

  b. tal como um carro que começa a diminuir a velocidade por falta de combustível

The annotation had to deal with clear cases of Identity but also with Near-Identity relations, predicative relations and bridging. Results are positive considering the complexity of the task, the lack of training of the annotators and the level of granularity of the guidelines that were distributed. Also, we believe that

the manual edition of the NPs would have made the task easier and provided conditions for a higher level of agreement among the annotators.

5 Conclusion

The main conclusion of our participation in this task is the high level of difficulty of the annotation process: it requires some computer science skills and a high linguistic knowledge. As a consequence of these requirements all the annotators without strong linguistic background failed to finish the task.

The use of pre-identified NPs is also a potential problem: if they are correct it helps the annotation process; but if they are incorrect this brings an overhead to the process having, as a consequence, the need to manually undo the initial annotation. The same problem is associated with the use of the visual editor: it helps when the NPs are correct but it showed to lack stronger editing options, allowing to easily change pre-identified segments.

In spite of all the described problems we believe this task allowed us to better understand the complexity and the details of coreference annotation and to contribute to the creation of a reference annotated corpus for the Portuguese language.

6 Acknowledgments

This work was partially supported by national funds through FCT – Fundação para a Ciência e Tecnologia, under projects PEst-OE/LIN/UI0214/2013 and PEst/CEC/UI04668/2013.

References

1. Bouma, G., Daelemans, W., Hendrickx, I., Hoste, V., Mineur, A.: The corea-project, manual for the annotation of coreference in dutch texts. Tech. rep. (2007)
2. Fonseca, E.B., Vieira, R., Vanin, A.: Corp: Coreference resolution for portuguese. In: 12th International Conference on the Computational Processing of Portuguese, Demo Session (PROPOR) (2016)
3. Fonseca, E.B., Sesti, V., Antonitsch, A., Vanin, A.A., Vieira, R.: Corp - uma abordagem baseada em regras e conhecimento semântico para a resolução de correferências. Linguamatica 9(1), 3–18 (2017)
4. Mendes, A.: Organização textual e articulação de orações. pp. 1691–1755. Gramática do Português, vol. II. Lisboa: Fundação Calouste Gulbenkian (2013)
5. do Nascimento, M.F.B., Mendes, A., Pereira, L.: Providing on-line access to portuguese language resources: corpora and lexicons. pp. 1825–1828. Proceedings of the IV International Conference on Language Resources and Evaluation - LREC2004, Lisbon, Centro de Cultural de Belém (2004)
6. Tubino, M.d.O., Silva, M.M.S.: Visualização, manipulação e refinamento de correferência em língua portuguesa. Trabalho de conclusão de curso, Pontifícia Universidade Católica do Rio Grande do Sul (2015)