From manual to automatic annotation of coreference

Renata Vieira*, **Caroline Gasperin***, **Rodrigo Goulart***
PIPCA - Unisinos
São Leopoldo, Brazil
{renata,caroline,rodrigo}@exatas.unisinos.br

Abstract

We present experiments on manual annotation of coreference in Portuguese texts motivated by the goal of developing and evaluating a coreference resolution tool. The tool architecture, designed to deal with multi-level annotation and to allow easy combination of heuristics and configuration of parameters, is also described in the paper.

1 Introduction

We are studying coreference in romance languages with the goal of developing a multi-lingual tool for coreference resolution. In this paper we present results in corpora annotation of coreference in Portuguese texts[1]. These experiments serve as background for the development of our tool. First, the tasks to be performed by the tool should be tasks that subjects can perform themselves and furthermore, there must be agreement in the analyses. Secondly, our experiments are developed to raise relevant features to be taken into account in the process of coreference resolution. The tool we are designing handles multi-level annotation encoded according to recently proposed standards.

In the next section, our experiments on Portuguese coreference annotation are detailed. In section 3, we present the designing principles of our tool. Conclusions of this work are presented in section 4.

[1]Both Brazilian and European Portuguese.

*Research Grant CNPq- Brazil.

2 Manual annotation of coreference

Coreference on natural language texts consists on two or more expressions referring to a same discourse entity. When they follow one another in the text we refer to the previous expression in the sequence as antecedent. We refer to terms under analysis in our studies as coreferent terms. Coreferent terms can be of different types, for example:

- pronominal: the coreferent term is a pronoun. Ex: *The boy read the book, but **he** didn't like it.*
- definite description: the coreferent term is a noun phrase preceded by a definite article. Ex: *I bought a house. **The house** is far from here.*
- demonstrative descriptions: the coreferent term is a noun phrase preceded by a demonstrative pronoun. Ex: *I bought a house. **This house** will be mine forever.*

In our work we have undertaken three experiments on manual annotation, considering two types of coreferent expressions (definite descriptions and demonstratives) and different annotation methodologies (as described later in the paper). We studied definite descriptions and demonstrative noun phrases separately, observing different features for each one. The annotation methodology has evolved by considering questions related to inter-annotator agreement. Table 1 presents an overview of the Portuguese corpora we studied.

The first two experiments presented next, were first presented in detail in (Salmon-Alt and Vieira, 2002; Vieira et al., 2002b; Vieira et al., 2002a).

Table 1: Language resources.

Experiment	Size (words)	Number of cases
Exp 1	5000	541 definites
Exp 2	50000	243 demonstratives
Exp 3	6795	730 definites

These previous works refer to both Portuguese and French languages.

We used MMAX (Müller and Strube, 2001) tool for manual annotation of coreference in the three experiments. MMAX was suitable for annotating the resources in a format as close as possible to the MATE recommendations, especially concerning the use of XML, the stand-off annotation principle and the compatibility with proposed coreference encoding guidelines (Poesio, 2000).

2.1 First experiment

The corpus used in the first experiment was composed of European Portuguese written question-answer pairs published in the Official Journal of the European Commission. Our classes of analyses were based on the analyses of English texts presented in (Poesio and Vieira, 1998), with the difference that we divided the *Bridging* class of their analyses into two different classes, separating coreferent (*Indirect Anaphora*) and non-coreferent (*Other Anaphora*) cases. The study aimed to verify if we could get a similar distribution of types of definite descriptions for Portuguese and English, which would serve as an indication that the same heuristics tested for English (Vieira and Poesio, 2000) could apply for Portuguese. The main annotation task in this experiment was identify antecedents and classify each definite description (d) into one of the following four classes:

- *Direct Coreference* d corefers with a previous expression a; d and a have the same nominal head:

  a. A Comissão tem conhecimento **do livro**... (the Commission knows *the book*)

  d. a Comissão constata ainda que **o livro** não se debruça sobre a actividade das várias ... (the Commission remarks that *the book* ignores the activity of various)

- *Indirect Coreference* d corefers with a previous expression a; d and a have different nominal heads:

  a. a circulação **dos cidadões** que dirigem-se (...) (the flow of *the citizens* heading to...)

  d. do controle **das pessoas** nas fronteiras (the control of *the people* in the borders)

- *Other Anaphora* d does not corefer with a previous expression a, but depends for its interpretation on a:

  a. **o recrutamento de pessoal científico e técnico**... (*the recruitment of scientific and technical employees*)

  d. **as condições de acesso à carreira científica** (*the conditions of employment for scientific jobs*)

- *Discourse New* the interpretation of d does not depend on any previous expression:

  d. o livro não se debruça sobre **a actividade das várias organizações internacionais**... (the book ignores *the activity of various international organisation*...)

2.1.1 Results

Table 2 presents the distribution of the annotation among the classes presented above. The distribution of uses is similar to previous studies for English texts.

Around 40% of the cases are discourse new, for these cases resolution does not apply. This indicates that the heuristics to identify discourse new descriptions proposed for English should be tested for Portuguese. These heuristics are mainly based on syntactic complexity. The syntactic structure of discourse new descriptions was analyzed for Portuguese. It was found that they were modified (by adjectives, prepositional phrases and relative

Table 2: Experiment 1.

Classification	Ann 1	Ann 2	Average
Direct coreference	96	179	25.4
Indirect coreference	51	45	8.9
Other anaphora	46	77	11.4
Discourse new	266	198	42.9
Not classified	82	42	11.5
TOTAL	541	541	100.0

clauses) in approximately 50% of the cases. These findings indicate that the heuristics developed for English may be applied to Portuguese.

Those classes that rely on common sense knowledge (indirect and other anaphora), and are, therefore, of difficult computational treatment, account for 20% of the total.

The agreement (given by Kappa) among the annotators was low, K = 0.44. This was worse agreement than the experiments made with English corpus. This could be related to the inclusion of a fourth class in the analysis (the splitting of bridging into indirect coreference and other anaphora). The motivation for introducing this class was to distinguish coreferent from non-coreferent (associative) uses like in *house - the door*. At first we considered that a better specification of the classes could improve the agreement results, but the results lead us to conclude that it could be more difficult to the annotators to deal with a greater number of choices.

2.2 Second experiment

In this experiment we analysed demonstrative noun phrases. Our classes here, similar to those used for definite descriptions, serve to estimate the frequency of antecedents that are noun phrases, the frequency of coreferential and other relations between demonstratives and their NP antecedents. We were also verifying the frequency in which the NP antecedent has the same head noun of the demonstrative. The reason why we are isolating nominal antecedents from other expressions such as verb phrases, sentences or paragraphs is that this can give us an idea of how well a system for coreference resolution of demonstratives can perform on the basis of nominal expression relations only.To gather this sort of knowledge, each demonstrative description (d) was classified into one of the following classes:

- *Direct coreference*, as before:

  a. e prestar **às autoridades gregas** (*to the greek authorities*)

  d. para **essas autoridades** (*these authorities*)

- *Indirect coreference*, as before:

  a. **À Albânia**

  d. ajudar **este país** a atingir (*this country*)

- *Other kind of anaphora* the antecedent is not a nominal expression or the relation between demonstrative and its antecedent is not a coreference relation.

  a. **o ano de 1993 será essencialmente consagrado ao apoio de experiências-piloto de informação dos jovens na Europa** (*the year of 1993 will be important to the experiments ...*)

  d. **nesse contexto** *in this context*

  a. **adoptar medidas de âmbito nacional** (*to adopt measures...*)

  d. **essa adopção** *this adoption*

2.2.1 Results

Table 3 shows the results obtained in the second experiment. The results show that demonstratives are context dependent, with nearly half of them being coreferent to previous NPs. The other half is either coreferent with antecedents that are not NP or not coreferent. Demonstratives whose antecedents were not explicitly marked by the annotator were included in other anaphora class; most of these cases were antecedents corresponding to more than one paragraph in the text.

Table 3: Experiment 2.

Classification	Ann 1	Ann 2	Average
Direct coreference	80	74	31.7
Indirect coreference	60	49	22.4
Other anaphora	77	66	29.4
Discourse new	0	0	0
Not marked	26	54	16.5
TOTAL	243	243	100.0

We calculated Kappa for three classes (direct coreference, indirect coreference, other). We found K = 0.65 for Portuguese demonstratives. These results show better agreement than for previous experiments related to four different classes for definite descriptions. The improvement might be related to the reduced number of classes and the kind of distinction involved.

2.3 Third experiment

In the third experiment we analysed again definite descriptions but adopting a different annotation methodology. The change in annotation methodology is related to the low agreement observed in the first experiment with definite descriptions. This could be due to some difficulties in the annotation process. There, all the annotation had to be done in a single step. In order to simplify annotators' tasks, we decided to split the annotation process in 4 steps (the first on is done by just one annotator and the others by two):

1. selecting coreferent terms;

2. identifying the antecedent of coreferent terms selected in sep 1, if there is one;

3. classifying coreferent terms: if there is an antecedent and it is coreferent, their relation should be classified (direct or indirect);

4. classifying non coreferent terms: if the expression doesn't have an antecedent or if it has a non coreferent antecedent classify the non coreferent relation (discourse new or other anaphora).

2.3.1 Results

The results here show the similar distribution of the previous experiments for definite descriptions. Compared to the total (730 cases) we have about half of them classified as discourse new descriptions, which account for about 70% of non-coreferent cases. Among the coreferent cases the number of direct coreference is twice the number of indirect coreference. Regarding the agreement among annotators, we see that after dividing our experiments in steps, the agreement of our annotators increased a little compared to experiment 1. Considering 4 classes (direct, indirect, discourse new, and other anaphora) we have K = 0.52. Considering Kappa for each step in the annotation task we have for step 2 (coreferent X non coreferent) K = 0.76, for step 3 (direct X indirect) K = 0.57 and for step 4 (other anaphora X dicourse new) K = 0.29. Clearly, the difficult class to analyse is the non coreferent class, that is the distinction between those cases introduced in text by an associate entity and cases based on subject's previous world knowledge. This confirms previous work done for English. Regarding the development of a tool for coreference resolution in texts, we can only hope to be able to identify coreferent from non-coferent terms, since this is the task that can be performed by speakers.

3 Automatic annotation of coreference

Based on the corpus studies presented in the previous section, we are developing a tool that is able to identify automatically the antecedents of coreferent expressions. The tool is designed on the basis of standard encoding of linguistics resources (Ide and Romary, 2002). It deals with multi-level annotated resources combining POS, syntactic and coreference

Table 4: Experiment 3.

	All cases	Ann 1	Ann 2	Average
Step 2	Coreferent	218	218	29.9
	Non coreferent	512	508	69.9
	None	0	4	0.2
	TOTAL	730	730	100.0
Step 3	Coreferent	Ann 1	Ann 2	Average
	Direct	125	151	63.3
	Indirect	93	67	36.7
	Other	0	0	0.0
	None	0	0	0.0
	TOTAL	218	218	100.0
Step 4	Non coreferent	Ann 1	Ann 2	Average
	Discourse new	354	382	72.2
	Other anaphora	147	114	25.6
	Other	11	12	2.2
	None	0	0	0.0
	TOTAL	512	508	100.0

```
 ...
<word id="word_92">as</word>
<word id="word_93">autoridades</word>
<word id="word_94">gregas</word>
 ...
<word id="word_135">essas</word>
<word id="word_136">autoridades</word>
 ...
```

Figure 1: Words file

information.

3.1 Input Data Format

Input data for our tool follows MMAX´s word and markable file formats. MMAX formats are shown on Figure 1.

The coreference annotation produced by manual analysis is encoded as shown on Figure 2, where the attribute *span* indicates the words that form each <markable>, the attribute *pointer* indicates the antecedent identifier, and the attribute *classification* corresponds to the classes presented earlier.

We also produce compatible POS and syntactic information[2] used for coreference resolution from the Portuguese parser PALAVRAS (Bick, 2000). PALAVRAS output is converted into a set of XML

[2]Presented in detail in (Gasperin et al., 2003)

```
 ...
<markable id="markable_3"
          span="word_92..word_94"
          pointer=""
          np_form="defNP"
          classification=""/>
 ...
<markable id="markable_5"
          span="word_135..word_136"
          pointer="markable_3"
          np_form="demNP"
          classification="indirect"/>
 ...
```

Figure 2: Markables file

```
 ...
<word  id="word_92">
  <art canon="o"
       gender="F"
       number="P">
    <secondary_art tag="artd"/>
  </art>
</word>
 ...
<word  id="word_93">
  <n   canon="autoridade"
       gender="F"
       number="P">
</word>
 ...
```

Figure 3: POS file

```
...
<chunk id="chunk_15" ext="np" span="word_92..word_94">
  <chunk id="chunk_16" ext="h" span="word_93"/>     </chunk>
...
<chunk id="chunk_27" ext="np" span="word_135..word_136">
   <chunk id="chunk_28" ext="h" span="word_136"/>
</chunk>
...
```

Figure 4: Chunks file

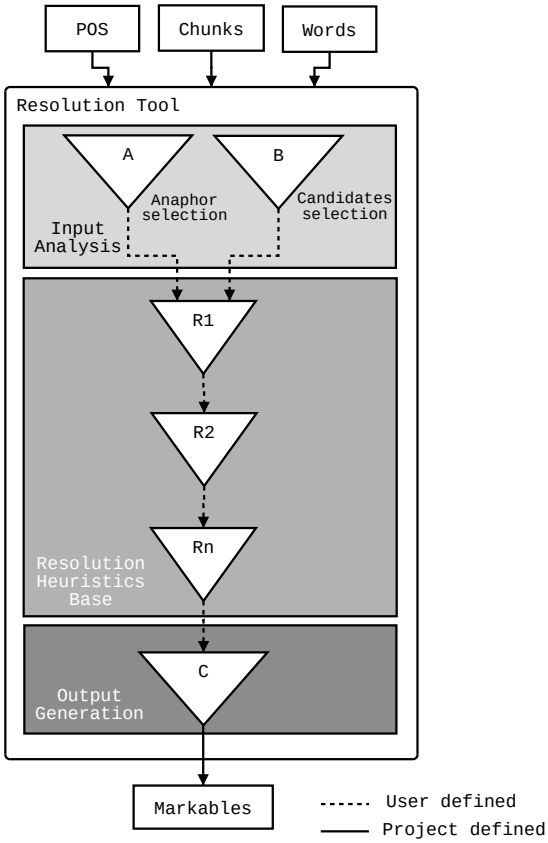

Figure 5: General architecture

files: the words file; a file with the part-of-speech (POS) categories for each word, an example is given in Figure 3, where the element *n* indicates a noun; and a file with the sentences syntactic structure, represented by <chunk> elements. A <chunk> represents a structure inside the sentence that can contain other child chunks, an example is shown in Figure 4, where the parent <chunk> (attribute @id =”chunk_7”) is a noun phrase (attribute ext=”np”) and the child <chunk> 8 is the header of the parent <chunk> (attribute ext=”h”).

3.2 General arquitecture

The design of the tool is based on “Pipes & Filters” proposed by Gamma (Gamma et al., 1995). The tool is composed by a set of filters that transform the data flowing through it. Each filter corresponds to a XSL script (which are simple and flexible) and implements heuristics for coreference resolution or parsers input and output of elements. The filters are combined into layers as shown in Figure 5.The first layer, “Input Analysis”, uses input data to generate two new data elements <anaphor> and <candidate> (Figure 6 and 7).

The <anaphor> elements are expressions to be resolved (definite, demonstratives or other noun phrases), which are extracted from corpus. This extraction is made by looking for certain attribute values in <chunk> elements. The presence of the definite articles in NPs, for instance, identifies definite descriptions.

The <candidate> elements are antecedent candidates for coreference resolution, different choices may also apply according to the heuristics adopted. We have considered NPs as candidates, for instance.

These new elements add one annotation level to our resources. They are used together with the other annotation levels in the layer called “Resolution Heuristics Base” (RHB). RHB layer is com-

```
...
<anaphor span="word_92..word_94"
         pointer=""/>
...
<anaphor span="word_135..word_136"
         pointer="word_92..word_94"/>
...
```

Figure 6: Anaphor elements

```
...
<candidate span="word_92..word_94"/>
...
<candidate span="word_135..word_136"/>
...
```

Figure 7: Candidate elements

posed by a set of filters corresponding to coreference resolution heuristics. The user of the tool can define the combination of heuristics and parameters to be taken into account at each execution.

Some resolution heuristics may be based on the comparison of the head nouns of <anaphor> and <candidate>. Their span values point to the corresponding <chunk> nodes. The heads are the childs of these chunks, having ext attribute value equal to "h" (chunk_16 and chunk_28). They are accessible through words and POS files ("autoridades"). When a suitable antecedent is found the pointer attribute of the <anaphor> will take the span attribute from the matching <candidate>. Other heuristics may combine information from POS and Chunks regarding the candidates and anaphors. The use of linguistic information changes according to the rules to be applied by the tool, as specified by the user.

The last layer, “Output Generation”, takes RHB´s output and adapts it according to the required format. In our case we generate MMAX markables, so we can evaluate our output with manually annotated corpus and also visualize the coreference chains in MMAX. However, other formats can be generated, for instance, the virtual annotation language proposed in (Ide and Romary, 2003; Ide and Romary, 2001). The combination of heuristics (different rules, different sequences and paramenters) may be defined and tested, since the user defines conections of RHB filters.

4 Conclusion

We have presented our work towards the designing of a coreference annotation tool on the basis of manual annotation experiments. Our experiments show that definite descriptions are commonly used to introduce new discourse elements in Portuguese texts, confirming previous findings for English. We also compared the use of definite descriptions to another type of noun phrases, commonly considered as anaphoric: demonstrative noun phrases. As opposite to definite descriptions, they are mainly text dependent for their interpretation (coreferent or anaphoric); antecedent NPs account for at least 50% of the cases (direct and indirect coreference). Therefore, the same heuristics applied to direct coreference of definite descriptions may be used for demonstratives, by the distribution we have an idea that the heuristics may account for 30% of the cases.

Following our corpus studies, we presented the general architecture of a tool for automatic coreference resolution. This tool process parsed corpus encoded in XML according to recommendations of the standards under development for corpora annotation (XCES (Ide and Romary, 2002), ISO TC37 SC4). The advantage of having data encoded in XML is the possibility of using the existing tools for handling XML data, as well as existing tools for manual coreference annotation (MMAX).

Regarding the tasks to be done automatically by the tool, we will concentrate in distinguishing between coreferent from non-coferent terms, and resolve co-referent terms, we will not try to resolve other anaphora, since we will not be able to evaluate

this task.

Although we concentrated this paper on the study of Portuguese text, the tool is being conceived to treat multilingual corpora, and we are first considering romance languages. Studies on French corpora (Salmon-Alt and Vieira, 2002; Vieira et al., 2002b; Vieira et al., 2002a), is being conducted along with our experiments for Portuguese.

Acknowledgements

We are grateful to CNPq, INRIA and FAPERGS for financial support. We would like to thank the help of Eckhard Bick, Christoph Müller, Michael Strube, Paulo Quaresma and Susanne Salmon-Alt.

References

Eckhard Bick. 2000. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, Århus University, Århus.

Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. 1995. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley Professional Computing Series. Addison-Wesley Publishing Company, New York.

Caroline Gasperin, Renata Vieira, Rodrigo Goulart, and Paulo Quaresma. 2003. Extracting xml syntactic chunks from portuguese corpora. In *Traitement automatique des langues minoritaires - TALN 2003*, Btaz-sur-mer, France.

Nancy Ide and Laurent Romary. 2001. Common framework for syntactic annotation. In *Proceedings of ACL'2001*, pages 298–305, Toulouse.

Nancy Ide and Laurent Romary. 2002. Standards for language resources. In *Proceedings of the LREC 2002*, pages 839–844, Las Palmas de Gran Canaria.

Nancy Ide and Laurent Romary. 2003. Encoding syntactic annotation. In Anne Abeillé, editor, *Building and Using Syntactically Annotated Corpora (in press)*. Kluwer, Dordrecht.

Christoph Müller and Michael Strube. 2001. MMAX: A tool for the annotation of multi-modal corpora. In *Proceedings of the IJCAI 2001*, pages 45–50, Seattle.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Massimo Poesio. 2000. Coreference.mate dialogue annotation-deliverable d2.1. Technical report, http://www.ims.uni.stuttgart.de/projekte/mate/mdag, Jan.

Susanne Salmon-Alt and Renata Vieira. 2002. Nominal expressions in multilingual corpora: Definites and demonstratives. In *Proceedings of the LREC 2002*, Las Palmas de Gran Canaria.

Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):525–579.

Renata Vieira, Susanne Salmon-Alt, Caroline Gasperin, Emmanuel Schang, and Gabriel Othero. 2002a. Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. In *Proceedings of the DAARC 2002*, Estoril.

Renata Vieira, Susanne Salmon-Alt, and Emmanuel Schang. 2002b. Multilingual corpora annotation for processing definite descriptions. In *Proceedings of the PorTAL 2002*, Faro.