From concrete to virtual annotation mark-up language: the case of COMMOn-REFs

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

Susanne Salmon-Alt
ATILF-CNRS
Nancy, France
Susanne.Alt@loria.fr

Abstract

This work presents the data model we adopted for annotating coreference. Our data model includes different levels of annotation, such as part-of-speech, syntax and discourse. We compare our encoding schemes to the abstract XML encoding being proposed as standard. We also present our tool for coreference resolution that handles our data model.

1 Introduction

We have been dealing with corpus based studies since 1997 (Renata Vieira and Simone Teufel, 1997; Poesio et al., 1997). Our focus has been the study of coreference. In the study of coreference we have dealt with annotation experiments (manual and automatic) and their respective annotation schemes. To work on coreference we used information from syntactic annotated corpus, the Penn Treebank. Our results (annotated corpus with coreference links and classification of coreference status) were Prolog encoded. When we first adapted our tool for Portuguese (Rossi et al., 2001) we dealt with other tools and annotation formats. The resources built on these previous works were difficult to share due to their particular information encoding.

Our current work in the COMMOn-REFs project (A computational model for processing referring expressions)[1], involves Portuguese and French. We are using MMAX, a tool for multimodal annotation in XML (Müller and Strube, 2001), for manual annotation of coreference, and we are developing a tool for automatic coreference resolution. Our tool deals with XML encoding provided by MMAX and syntactic information for Portuguese and French encoded in XML. In order to be able to share the resources being built, we are relating our model with proposed standards.

In Section 2 we present previous annotation formats that we dealt with. In Section 3 we give an overview of the work in COMMOn-REFs. Section 4 relates our current model with the standards recently proposed (Ide and Romary, 2002; Ide and Romary, 2003; Ide and Romary, 2001). Section 5 describes our tool for coreference resolution. A discussion on the problems we face with our annotation model is presented in Section 6.

2 Previous work

Our first annotation schemes were Prolog lists of treebank sentences and their noun phrases (NPs), as shown in Figure 1. The lists were extracted from Lisp lists of the Penn Treebank. These lists were manipulated in our experiments on coreference annotation and resolution.

The results of coreference annotation were lists of Prolog facts `dcc(Index1,Index2,Code)` as shown in Figure 2. `Index1` refers to the sequential numbering of definite descriptions; `Index2` refers to the sequential numbering of noun phrases; and `Code` refers to their classification, according to discourse status (Poesio and Vieira, 1998). For some of them there were also facts

[1]http://www.inf.unisinos.br/ renata/documents/commonrefs-proposal.pdf.

* Research Grant CNPq- Brazil.

```
[S,[NP,the,squabbling,[PP,within,[NP,the,
Organization,[PP,of,[NP,Petroleum,Exporting,
Countries]]]]],[VP,seems,[PP,under,
[NP,control]],[PP,for,now]].].

[NP,Petroleum,Exporting,Countries].

[NP,the,Organization,
[PP,of,[NP,Petroleum,Exporting,Countries]]].

[NP,the,squabbling,[PP,within,
[NP,the,Organization...
```

Figure 1: Prolog NP and S lists.

```
...
ddc(5, 16, r).
ddc(6, 18, k).
ddc(7, 28, r).
...
ddsr(5, 16, r, np(5)).
ddsr(7, 28, r, np(16)).
...
```

Figure 2: Prolog coreference annotation.

`ddsr(Index1,Index2,Code,Antecedent)` indicating their antecedent NPs. We could only link the annotation to the data by running the Prolog code that loaded the lists of NPs and sentences and generated their indexes. Although we had carried out intensive research with these resources and tools, the re-use of our data in other environments was very difficult.

Despite the lack of fully annotated data for Portuguese, we tried to check out whether the same heuristics we used for English would be suitable for this new language. To test our heuristics we used the PALAVRAS parser[2] (Bick, 2000) to parse Portuguese corpus. From parsed texts we extracted Prolog lists of NPs as illustrated in Figure 3. Experiments were carried out over these resources. Heuristics for correference resolution were adapted to Portuguese and the results obtained were comparable to those previously obtained for English. However, the genericity of the Portuguese resolver and annotated data still raised the same re-usability problems as for English, since the encoding format had not evolved.

3 COMMOn-REFs

In the COMMOn-REFs project we face the challenge of dealing with different languages (French and Portuguese). Therefore, we have to share corpora and tools, initially available under different formats.

```
STA:fcl
P:v-fin('ser' PR 3P IND) São
SC:adj('remoto' F P) remotas
SUBJ:np
=>N:art('o' <artd> F P) as
=H:n('chance' F P) chances
=N<:pp
==H:prp('de') de
==P<:np
===H:n('aprovação' F S) aprovação
....

[NP,as,[N,chances],[PP,de,
[NP,[N,aprovacao]]]]

[NP,[N,aprovacao]]
```

Figure 3: PALAVRAS output.

We adopted MMAX[3] as our manual annotation tool. With MMAX we could annotate our corpus according to our theoretical principles. The following corpus studies were developed with the aid of the tool: (Salmon-Alt and Vieira, 2002; Vieira et al., 2002b; Vieira et al., 2002a). In these studies, our annotation targets were manually marked and coreference information was added to them according to subjects' analysis of the texts.

We are currently developing a coreference resolution tool on the basis of XML files and XSL scripts. The tool manipulates several levels of linguistic information. Parsing information has been provided by the PALAVRAS parser. The parser output is transformed into two XML files: one with POS and another with syntactic information (chunks). Coreference information, manually annotated with MMAX (markables), is used for evaluation. Our tool, besides manipulating three different annotation levels (POS, chunks, markables), creates two others: anaphors and candidates, as detailed in Section 5.

As we are interested in having our resources made available, we relate our annotation schemes to standard proposals presented in (Ide and Romary, 2002; Ide and Romary, 2003; Ide and Romary, 2001).

[2] http://visl.hum.sdu.dk/visl/pt

[3] The MMAX tool is available for download in http://www.eml.org/english/Research/NLP/Downloads.

```
<markables>
...
  <markable id="markable_1"
    coref="no"
    classif="disc_stat_1"
    span="word_1..word_2" />
...
  <markable id="markable_3"
    coref="yes"
    pointer="markable_1"
    classif="disc_stat_2"
    span="word_8" />
...
</markables>
```

(a)

```
<struct type="CRAnnot">
...
  <struct id="m1" type="C-level">
    <feat type="coref">no</feat>
    <feat type="classif">disc_stat_1</feat>
    <feat type="target">#w1 #w3</feat>
  </struct>
  <struct id="m3" type="C-level">
    <feat type="coref">yes</feat>
    <feat type="pointer">#m1</feat>
    <feat type="classif">disc_stat_2</feat>
    <feat type="target">#w8</feat>
  </struct>
...
</struct>
```

(b)

Figure 5: Markables file.

```
<words>
  <word id="word_1">O</word>
  <word id="word_2">jogador</word>
  <word id="word_3">pode</word>
  <word id="word_4">deixar</word>
  <word id="word_5">o</word>
  <word id="word_6">time</word>
  <word id="word_7">.</word>
  <word id="word_8">Ele</word>
  <word id="word_9">recebeu</word>
  <word id="word_10">uma</word>
  <word id="word_11">proposta</word>
  <word id="word_10">excelente</word>
  <word id="word_11">.</word>
</words>
```

Figure 4: Words file.

4 Data model

4.1 Encoding standards

Directions for standard corpus encoding in XML have been proposed in (Ide and Romary, 2002; Ide and Romary, 2003; Ide and Romary, 2001). Such efforts consist on defining abstract formats for corpus annotation that could be instantiated according to specific project requirements. An abstract XML file can be generated for each annotation level according to a Virtual Annotation Markup Language (VAML). The structure of this language is defined by a skeleton that consists on <struct> (a node/level in the annotation) and <feat> elements (feature attached to the enclosing <struct> node).

Particular project specifications are defined through data categories (component categories to be annotated) and dialect (encoding style). On the basis of these specifications, a mapping between VAML and Concrete AML (CAML) can be made. CAML is the language used for annotation encoding in particular projects.

4.2 Our schemes

Our first experiments with MMAX were on manual coreference annotation. The tool required specific input and output formats. Our corpus, that is, the primary data, were first converted from raw texts to XML, encoded as <word> elements, like the example in Figure 4 for the sentences *O jogador pode deixar o time. Ele recebeu uma proposta excelente. (The player may leave the club. He received an excellent proposal.)* Each corpus token (words and punctuation) corresponds to a <word>.

The coreference was manually annotated and encoded as <markable> elements. Each anaphoric expression and antecedent were represented by markables. Anaphors' markables had an extra attribute "pointer", that refers to its antecedent markable. An example of a markable file is presented on Figure 5(a). Markables correspond to our final level of annotation. The "span" attribute refers to our primary data, the words. The other attributes (coref, classific) were specified according to our application. Figure 5(b) represents the abstract XML encoding for our markables file, according to VAML[4]. Pointer, coreference and classification compose our dialect vocabulary for the following data categories: antecedent and types of discourse status, as in (Poesio

[4]As we are not aware of a registry of data categories for coreference level, in our examples throughout the paper we often use the same vocabulary in abstract and concrete encodings.

```
<words>
  <word id="word_1">
    <art canon="o" gender="M" number="S"/>
  </word>
  <word id="word_2">
    <n canon="jogador" gender="M" number="S"/>
  </word>
...
</words>
```

(a)

```
<struct type="MSAnnot">
  <struct id="w1" type="W-level">
    <feat type="pos">ART</feat>
    <feat type="lemma">o</feat>
    <feat type="gender">M</feat>
    <feat type="number">S</feat>
    <feat type="target">#w1</feat>
  </struct>
  <struct id="w2" type="W-level">
    <feat type="pos">N</feat>
    <feat type="lemma">jogador</feat>
    <feat type="gender">M</feat>
    <feat type="number">S</feat>
    <feat type="target">#w2</feat>
  </struct>
...
</struct>
```

(b)

Figure 6: POS file.

and Vieira, 1998) inpired on (Prince, 1981; Prince, 1992). According to our dialect instantiation style, the data categories are represented as attributes of <markable> elements.

To develop a tool for automatic coreference resolution, we needed to consider other intermediate levels of annotation: source of linguistic information used for solving anaphoras.

Our corpus was analysed by the Portuguese parser PALAVRAS (Bick, 2000). The original format of PALAVRAS output is not standard. As previously presented in Figure 3, on each line of the figure:

- the first symbol represents the syntactic function (‘SUBJ’= subject, ‘N’=noun modifier, ‘H’=head, etc.);
- following ‘:’ , we have the syntactic form for groups of words (’np’=noun phrase, etc.) and POS-tags for single words (’n’=noun, ’v’=verb, etc.);
- in brackets are the word canonical form and other inflectional tags;
- after the brackets comes the word as it occurs in the corpus.

The ‘=’ signs in the beginning of each line represent the level of the phrase in the parsing tree[5].

[5]A complete description of the tagset symbols is available at http://visl.hum.sdu.dk/visl/pt/info/symbolset-manual.html.

We defined the XML encoding for PALAVRAS output. We split PALAVRAS output into two annotation levels, one for POS and another for syntactic data. Figure 6(a) shows our scheme for POS file. The corresponding abstract XML file is presented on Figure 6(b). Our data categories are word canonic form (lemma), pos, gender, number, person, tense, mode, and case. According to our dialect instantiation style, each POS data category is represented by a new XML element, the other inflexional tags are encoded as attributes of this element. By handling a parsed corpus we could treat compounds at word level; the multi word expression “São Paulo”, for example, is tokenised as one word and codified as <word id="word_9">São_Paulo</word>.

We encode syntactic data as chunks. Each syntactic structure is represented by a <chunk> element. Figure 7(a) shows our encoding. The mapping to abstract XML is presented on Figure 7(b). In our dialect, each <chunk> in the concrete XML encoding corresponds to a <struct> in the abstract one.

5 Automatic coreference resolution

The tool we are developing for anaphora resolution takes as input word, POS and chunk files (the architecture design is shown on Figure 8). The resolution process is perfomed by a set of stylesheets, each one representing a different heuristic. This set is called Resolution Heuristics Base (RHB). A stylesheet is connected to another through pipes and it filters the information flowing through the system (Gamma et

```
<text>
  <paragraph id="paragraph_1">

    <sentence id="sentence_1" span="word_1..word_7">
      <chunk id="chunk_1" function="subj" form="np" span="word_1..word_2">
        <chunk id="chunk_2" function="n" form="art" span="word_1"/>
        <chunk id="chunk_3" function="h" form="n" span="word_2"/>
      </chunk>
      <chunk id="chunk_4" function="p" form="vp" span="word_3..word_4">
        <chunk id="chunk_5" function="aux" form="v" span="word_3"/>
        <chunk id="chunk_6" function="h" form="v" span="word_4"/>
      </chunk>
      <chunk id="chunk_7" function="acc" form="np" span="word_5..word_6">
        <chunk id="chunk_8" function="n" form="art" span="word_5"/>
        <chunk id="chunk_9" function="h" form="n" span="word_6"/>
      </chunk>
    </sentence>
...
</text>
```

(a)

```
<struct id="s0" type="T">
  <struct id="s1" type="P">
    <struct id="s2" type="S" span="word_1..word_7">
      <struct id="s3" type="NP" rel="subj" ref="word_1..word_2">
        <struct id="s4" type="art" rel="n-mod" ref="word_1"/>
        <struct id="s5" type="n" rel="h" ref="word_2"/>
      </struct>
      <struct id="s6" type="VP" rel="p" ref="word_3..word_4">
        <struct id="s7" type="v" rel="aux" ref="word_3"/>
        <struct id="s8" type="v" rel="h" ref="word_4"/>
      </struct>
      <struct id="s9" type="NP" rel="acc" ref="word_5..word_6">
        <struct id="s10" type="art" rel="n-mod" ref="word_5"/>
        <struct id="s11" type="n" rel="h" ref="word_6"/>
      </struct>
    </struct>
...
</struct>
```

(b)

Figure 7: Chunks file.

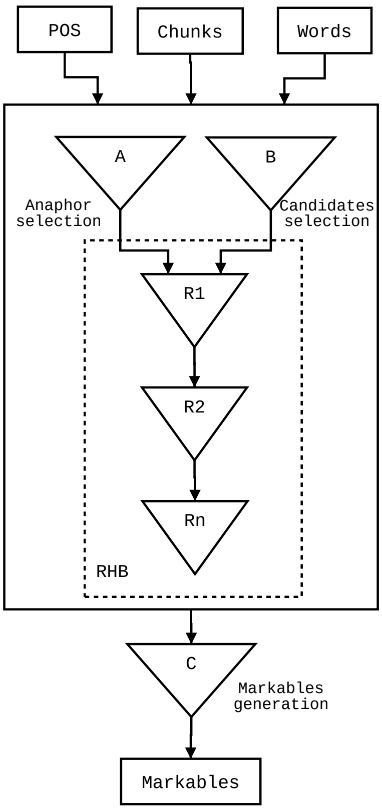

Figure 8: Anaphora resolution design.

al., 1995), and all heuristics can access the input files when necessary. Our tool strategy follows four main steps: anaphor selection, candidates selection, resolution, and output generation.

Two new intermediate annotation levels are generated: the anaphor entities (represented by <anaphor> elements) and antecedent candidates (represented by <candidate> elements).

The <candidate> represents possible antecedents in the corpus, and it also has a "span" attribute. (Figure 9(a)). Different candidate sets can be generated according to the heuristics used for its selection. demonstrates the corresponding VAML encoding. The <anaphor> depicts the anaphoric noun phrases (pronouns, definite descriptions, demonstratives) and it has the attribute"span" (Figure 9(b)). Through "span" value we can get information from the input files (words, POS, chunks), needed for the resolution process.

Along the resolution process other attributes are added to anaphor elements, such as "coref", "pointer" and "classif" attributes, as seen in Figure 10(a). Figure 10(b) represents the corresponding VAML encoding for the <anaphor> elements.

The heuristics to be applied to resolve coreference are based on previous studies about resolution of referring expressions (Vieira and Poesio, 2000; Lappin and Leass, 1994; Strube et al., 2002) and they are not discussed here.

The output is the last step in the process and it is also played by a stylesheet that translates the <anaphor> nodes into <markable> ones, so the results can be visualized using the MMAX tool.

6 Discussion

We have presented the evolution of our annotation schemes over 7 years of corpus research. We believe that a standard orientation may shed some light to those who are defining their projects. Concerning annotation level relations our annotation is based on object-based anchoring, especially because our primary data is represented by XML elements (words in our dialect, basic struct elements with id attributes in VAML).

Considering relations like parallelism, alternatives and aggregation (Ide and Romary, 2002) we see that our model includes aggregation at the chunk level. When studying annotation agreement we need to represent alternative data according to the judgment of each annotator (although we have adopted duplicated annotated files previously in our project).

Previous work on encoding standards has mentioned mainly POS and syntactic annotation. In this paper we extended its use for coreference annotation. Our data model could be adequately mapped to the standards.

An issue raised by coreference annotation is the need of two references for primary data in the same <struct>, one for anaphor (target) and another for its antecedent. In our examples, we encoded the reference to primary data indicating the antecedent by <feat> elements with attribute type="pointer".

An advantage we could expect to take from work related to standards is knowledge about the impact on performance in data handling according to encoding decisions.

Our project deals with different input and output formats. We intend to share our results and compare our techniques to different ones for anaphora reso-

```
<candidates>
  <candidate span="word_1..word_2"/>
  <candidate span="word_5..word_6"/>
  <candidate span="word_8"/>
</candidates>
```

(a)

```
<anaphors>
  <anaphor span="word_1..word_2"/>
  <anaphor span="word_5..word_6" />
  <anaphor span="word_8" />
</anaphors>
```

(b)

Figure 9: Candidate and anaphors.

```
<anaphors>
  <anaphor span="word_1..word_2"/>
  <anaphor span="word_5..word_6" />
  <anaphor span="word_8"
           coref="yes"
           classif="disc_stat_2"
           pointer="word_1..word_2"/>
</anaphors>
```

(a)

```
<struct type="ANAnnot">
  <struct type="A-level">
    <feat type="target">#w1 #w2</feat>
  </struct>
  <struct type="A-level">
    <feat type ="target">#w5 #w6</feat>
  </struct>
  <struct type="A-level">
    <feat type="target">#w8</feat>
    <feat type="coref">yes</feat>
    <feat type="classif">disc_stat_2</feat>
    <feat type="pointer">#w1 #w2</feat>
  </struct>
</struct>
```

(b)

Figure 10: Resolved anaphors.

lution. Since we use XML for external and internal encoding, and there is a mapping between them and standard formats, such as VAML, we will be able to import and export the corresponding VAML for our CAML and share both our resources and tools.

7 Acknowledgments

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

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.

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.

Shalom Lappin and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4).

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, Renata Vieira, and Simone Teufel. 1997. Resolving bridging descriptions in unrestricted texts. In *Proceedings of the Practical, Robust, Anaphora Resolution for Unrestricted Texts, Workshop on Operational Factors*, Madrid.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press, New York.

Ellen F. Prince. 1992. The ZPG letter: subjects, definiteness, and information status. In S. Thompson and W. Mann, editors, *Discourse description: diverse* *analyses of a fund-raising text*, pages 295–325. John Benjamins, New York.

Renata Vieira and Simone Teufel. 1997. Towards resolution of bridging descriptions. In *Proceedings of the ACL-EACL'97 Joint Conference: 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid.

Daniela Rossi, Clarissa Pinheiro, Nara Feier, and Renata Vieira. 2001. Resolução de correferência em textos da língua portuguesa. *Revista Eletrônica de Iniciação Científica*, 1(2).

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.

Michael Strube, Stefan Rapp, and Christoph Müller. 2002. The influence of minimum edit distance on reference resolution. In *Proceedings of the EMNLP 2002*, Philadelphia.

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.