Towards a Reference Annotation Framework

Susanne Salmon-Alt*, Laurent Romary**

* ATILF – CNRS
44 avenue de la Libération, B.P. 30687
54063 Nancy Cedex, France
Susanne.Salmon-Alt@atilf.fr

** LORIA
Campus Scientifique, B.P. 239
54506 Vandoeuvre-lès-Nancy Cedex, France
Laurent.Romary@loria.fr

Abstract

This paper discusses the main characteristics of a possible unified framework for specifying annotation schemes dedicated to the task of reference identification and linking on linguistic corpora. Built upon the foundation principles of the Linguistic Annotation Framework, the model (RAF, Reference Annotation Framework) is based on the combination of a simple meta-model (expressing markables and links between them) and a selection of data categories representing the information actually attached to each component of the meta-model. Based on the observation of existing practices we show how this model can be used in a variety of practical and theoretical configurations.

1. Introduction

Reference annotation associates referring expressions – usually certain types of noun phrases and pronouns – with information that enables their interpretation (e.g., their possible antecedents). This kind of knowledge is required for a variety of language processing applications, including information extraction and retrieval, natural language understanding and generation, machine translation, and human-machine dialogue.

Reference annotation in a broad sense, covering coreference, anaphora and reference encoding, has been subject of substantial practical and theoretical work during the last decade (Chinchor & Hirschman, 1997, Poesio & Davies, 2000, Poesio 2000, van Deemter & Kibble 2000, Tutin & al. 2000, Salmon-Alt 2001, Müller & Strube 2001, Vieira & al. 2002). Among them, van Deemter & Kibble (2000) suggest basic principles for coherent coding procedures from a linguistic point of view, whereas Poesio (2000) and Salmon-Alt (2001) made an attempt of unifying existing practices from a representational point of view. Example (1) (Poesio, 2000) illustrates current practice in coreference annotation: a coreferential link of type *identity* holds between a source markable *orange juice* and a target markable *orange juice*:

(1)
```
When do we have <coref:de ID="de _01">
orange juice </coref:de> at Elmira? We have
<coref:de ID="de _02"> orange juice
</coref:de> at Elmira at 6 a.m.

<coref:link type="ident"
href="coref.xml#id(de_02)">
  <coref:anchor
   href="coref.xml#id(de_01)"/>
</coref:link>
```

This paper is concerned with this latter issue, by assuming that it is possible, and indeed necessary, to fix up current practices in the field as a future standard discussed under the auspices of ISO committee TC 37/SC 4 on Language Resource Management[1]. Indeed, it is assumed that by achieving an international consensus on such a standard, it should be possible in the near future to share annotated resources, but above all to identify generic tools for editing and manipulating such data.

Our objective is to build upon the basic principles of annotation scheme specification suggested in (Ide & Romary 2002). This previous work also provides a default simplified syntax (GMT, Generic Mapping Tool) allowing one to make blind dump of annotation information for archival and/or exchange purposes in the case no specific XML syntax is available.

After a short presentation of these principles (section 2) we present the meta-model that informs the main characteristics of our reference annotation framework and propose a core set of data categories that may be used to instantiate such a meta-model in a specific application (section 3).

2. The Linguistic Annotation Framework

The model for specifying and representing reference annotation schemes that we present here is based on the general principles of the Linguistic Annotation Framework, the premises of which, being an on-going project within ISO committee TC 37/SC 4, have been stated in (Ide & Romary, 2002; Ide & Romary 2004). The general principles have already been implemented in the specific case of the representation of terminological data in the context of the design of ISO standard 16642 (ISO 16642, 2003). Those principles consider a class of semi-structured documents that can be specified through the combination of, on the one hand, a meta-model that informs the general practices in organizing information in a given application domain, and, on the other hand, a selection of data categories (DCS), that characterizes the

[1] See http://www.tc37sc4.org

elementary information units that can be attached to the various components of the meta-model. Indeed, the components in the meta-model should be considered as elementary linguistic abstractions that reflect the granularity of the description intended by the meta-model. For instance, Figure 1, Figure 2 and Figure 3 represent a very simple component corresponding to the description of the flexion of a lexical unit, as could be used in a wider meta-model for lexical databases. This level has been simply decorated by three data categories describing the actual form of the flexion, together with the corresponding gender and number. The assumption is that additional information concerning the word (e.g. part of speech) is inherited when the flexion level occurs within a wider lexical structure. In the same way, additional data categories are of course needed to describe the flexions of other types of words such as verbs, etc.

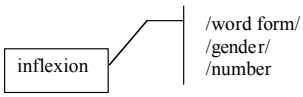

Figure 1: Simple combination of a meta-model level with data categories

```
<struct type="inflexion">
        <feat type="word form">vertes</feat>
        <feat type="gender">feminine</feat>
        <feat type="number">singular</feat>
</struct>
```

Figure 2 : GMT instance

```
<formeFléchie>
        <orth>vertes</orth>
        <genre>f</genre>
        <nombre>s</nombre>
</formeFléchie>
```

Figure 3: Ad hoc XML representation

As can be seen, one may derive a very simple representation format that matches isomorphically the model, as well as a specific XML structure, as long as the compatibility of the model is ensured. The LAF principles state that a specific linguistic annotation scheme can be described accordingly and assert some additional requirements on what it should necessarily contain and how it should be concretely implemented. Among them, we can quote here the equivalency between in-line and stand-off annotation, with the possibility of both inserting reference annotation mark-up directly into primary text data or separating primary data from annotation data by means of pointing mechanisms. Still, we consider stand-off markup as the reference model for primarily describing an annotation scheme.

3. From Current Practice to a Reference Annotation Framework

3.1. A meta-model for reference annotation

From the general principles of designing annotation schemes it is possible to derive a meta-model that covers the various features characterizing reference annotation. Figure 4 outlines the proposal of such a meta-model. The following sections describe the components of the meta-model and give a more closer view at data categories to be used for instantiating it.

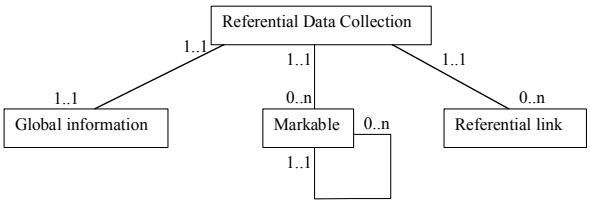

Figure 4: Meta-model for reference annotation

3.2. Components

The reference annotation scheme meta-model, organized around three main components, gathers up all information related to a specific annotation document within a global level named *Referential Data Collection*. Beside a *Global Information* component for the meta-data associated with the annotation file, it contains markables and referential links.

3.2.1. Markables

The basic constituents of any reference annotation scheme are, as an input, source markables, and, as an output, links to target markables[2]. Markables are either built upon parsed text chunks (noun phrases, pronouns etc.) or directly annotated on the source text. Depending on the underlying theory, they represent anaphoras and antecedents (Tutin & al, 2000), co-referring expressions (Chinchor & Hirschman, 1997) or referring expressions and referents (Bruneseaux & Romary, 1997).

In the framework presented here, *Markables* are the elementary units participating in anaphorical, coreferential or referential links. Markables may point to externalized source data (e.g. to words, morpho-syntactic units, syntactic chunks, representations for universe entities or gestures), from where relevant linguistic information (type of NP, gender, number, etc) may be percolated. However, they are autonomous – representing essential linguistic abstractions from source data – in two senses: First, they are not necessarily isomorphic to elements from the source data. This property is essential and allows for building complex markables recursively (e.g. for plural antecedents), for introducing relevant elements that are not present in any source data (e.g. zero pronouns) and for creating markables from row data (in this case, the source text is not a pointer, but a surface string). Second, markables may be characterized by features that are specific to the reference level (see section 3.3).

The following example shows an off-line representation for markables. (3) presents the primary source for the text in (2), supposedly segmented into word units, as for instance expected by an annotation software such as MMAX (Müller & Strube, 2001). Figure (4) shows the GMT representation of two markables with morpho-syntactic information which have percolated from lower levels.

[2] In practice, target markables are often supposed to be an input for the linking procedure. For a critical discussion of this practice, see van Deemter and Kibble (2000).

```
(2)   Prendre une poire et la faire cuire. Laver
      une pomme. Éplucher le fruit. Les faire
      glacer. Servir l'un chaud et l'autre frais.

(3)   <w id="w_1">une</w>
      <w id="w_2">poire</w>
      …
      <w id="w_4">la</w>
      …
      <w id="w_9">une</w>
      <w id="w_10">pomme</w>

(4)   <struct id="m_1"  type="markable">
        <feat type="source text"
        target="w_1..w_2"/>
        <feat type="syntactic category">
        noun phrase</feat>
        <feat type="determiner type">
        indefinite</feat>
        …
      </struct>

      <struct id="m_2"  type="markable">
        <feat type="source text"
        target="w_4"/>
        <feat type="syntactic category">
        pronominal phrase</feat>
        …
      </struct>
```

3.2.2. Referential Links

Any reference annotation schema makes use of (mostly typed) links between source and target markables. Those links represent a relation which has been considered by the annotator as necessary for correct discourse interpretation: depending on the theory, this could be an equivalence relation (coreferential links between expressions referring to the same entity are symmetrical, transitive and reflexive) or not (referential links from a referring expression to a referent or anaphorical links from an anaphor to an antecedent). However, current schemes can be distinguished on the basis of their use of an autonomous link element or not. Schemes using an autonomous link express relations between markables by means of a separate link element for the relation rather then just by a pointer attached to the source markable. An autonomous link element is however preferable for representing ambiguities and different links from the same source markable (Davies & Poesio, 2000).

Therefore, our reference annotation framework introduces a *Referential Link* component, relating markables that are linked by a specific referential relation. As will be seen in 3.3, this pointing mechanism is actuated by means of two data categories, */referential source/* and */referential target/*, that should be systematically part of any DCS derived from RAF. Referential links may also contain information about the type of the link. In example (5), the relation between the referents of the source markable *m_2* (*la*) and the target markable *m_1* (*une poire*) is encoded as an objectal relation of coreference.

```
(5)   <struct id="link_1" type="ref_link" >
        <feat type="objectal relation">
         coreference</feat>
        <feat type="ref source" target="m_2"/>
        <feat type="ref target" target="m_1"/>
      </struct>
```

3.3. Core Data Categories for RAF

This section discusses some issues related to the definition of core data categories related to reference, coreference and anaphora annotation. It concerns specific information to be attached to markables and links. Additionally to the feature discussed below, both markables and referential links can be associated with data categories used to indicate the origin (*/informer/*) and level of confidence (*/confidence level/* of the corresponding information).

3.3.1. Data Categories related to Markables

Beside relevant information that can be percolated from lower levels of annotation (*/grammatical gender/*, */grammatical number/*, etc.), markables must contain a data category */source text/* which identifies the underlying linguistic expression, either by means of a pointer to some external data or giving it explicitly. Furthermore, they may be associated with (a still open list of) semantic or referential information specific to the reference level:

Semantic information: Reference and anaphora resolution involves knowledge about the semantic properties of the underlying discourse entities. Therefore, annotators may wish to characterize markables further, for example by means of information about animacy, named entity categorization, word sense disambiguation, or more generally, entity types (based on an ontology).

Referential information: One theoretical issue in reference resolution is related to the referential status of the underlying discourse entities. Several authors proposed classifications (Hawkins 1978, Ariel 1990) that should be integrated in the data categories relevant for reference annotation. Another issue is the type of the expressions to be annotated. Annotators should be able to classify reference markables independently of morpho-syntactic information, for example for marking up different pronominal expressions or sub-types of expressions involved in temporal reference.

3.3.2. Data Categories related to Links

Referential Links necessarily have one */referential source/* data category, that is a pointer to the markable for which a link has to be found. They also have at most one */referential target/* feature, pointing to the markable to which the link has been established.

Furthermore, previous work on reference annotation has shown the need of typing the relation between the linked markables. However, as clearly pointed out by van Deemter & Kibble (2000), reference annotation in the sense considered here (covering coreference and anaphora) has to face the issue of properly characterizing the types of the relations to be covered. A comparison of types of relationships involved in current coreference annotation practice shows a very heterogeneous inventory (referential properties such as *identity of the referent*, set relations, semantic features such as linguistic bridging, *role in event*, function value relations, bound anaphora, etc.)[3]. On the other hand, it has been shown for several languages that acceptable inter-annotator agreement could

[3] see Poesio & Davies (2000) for an overview and van Deemter & Kibble (2000) for a critical analysis of MUC practice.

only be achieved on very basic distinctions (Poesio & Vieira, 1998).

As a conclusion for the design of RAF, we propose to introduce an explicit distinction between objectal and lexical relations. Objectal relations hold between the referents of the expressions to be annotated and include relations such as *coreference*, *part-of* or *set-subset* relations. Lexical relations hold between the expressions to be annotated and include *hypernomy*, *lexical identity*, *lexical bridging*. The definition of the list of values for each of these relations and their scope is still matter of discussion.

4. RAF in action : some complex cases

The basic principles sketched out in section 3 may also take into account the encoding of less straightforward configurations, which have often been considered as difficult ones. This is the case for plural antecedents, such as *les* referring to the set formed by *une poire* and *une pomme* (see example (2)). In RAF, the decision to use autonomous markables leads to the possibility of creating recursively complex markables, even for graphically disjoint surface sequences. The referential link holds then between a simple source markable and a complex target markable, as in (6):

```
(6)   <struct id="m_5" type="markable">
        <struct id="m_1" type="markable">
           <feat … </feat> …
        </struct>
        <struct id="m_3"  type="markable">
           <feat … </feat> …
        </struct>
      </struct>
```

Another complex case is the same source markable involved in several distinct anaphorical relations. For *l'autre* in (2), one could consider (and wish to annotate) a coreference link with *une pomme*, a *subset-of* link with *les* and a perhaps some theory-specific link with *l'un*. For those cases, RAF simply proposes to use as many as necessary distinct link structures, involving the same source markable, different target markables and different link types.

This case is still different from ambiguity, where several antecedents for a same source markable are mutually exclusive (see *le fruit* for which a system could hesitate between either *une pomme* or *une poire* as the right antecedent). For those cases, RAF recommends the use of the alternative structure <alt>, such as defined in (Ide & Romary 2004) and illustrated in (7):

```
(7)   <struct id="link_1" type="ref_link">
       <feat type="source text" target="m_4"/>
       <feat type="objectal link type">
        coreference</feat>
       <alt>
         <brack>
           <feat type="target" target="m_1"/>
         </brack>
         <brack>
           <feat type="target" target="m_3"/>
       </brack>
       </alt>
      </struct>
```

5. Conclusions and further work

The explicit statement of the underlying properties of reference annotation (especially the introduction of an autonomous markable and link component) as well as the ongoing discussion on relevant data categories) allows to localize several other issues, mentioned sometimes as being related to reference annotation, at more appropriate representation levels: disfluencies in oral discourse (*the…hum…dog*), zero pronouns (i.e. in Japanese), agglutinated markables (i.e. in romance languages) or ellipses are, for instance, rather a matter of morpho-syntactic representation whereas the integration of multi-modal reference (a pointing gesture to a discourse external object) into RAF should still be considered as an open issue. Another open issue is the definition of data categories for objectal and lexical relations, having in mind that the decision is not always straightforward. Some of the topics still under discussion are function-value relations, nominal predicates or bound anaphora.

6. References

Bruneseaux F., Romary L. (1997). Codage des références et coréférences dans les DHM. *Actes of ACH-ALLC '97*, Kingston Ont.

Chinchor N., Hirschmann L. (1997). MUC-7 Coreference Task definition, Version 3.0, *Actes de MUC-7*, 1997, http://www.muc.saic.com.

Davies S., Poesio M. (2000). MATE Deliverable 1.1, Chapter 3 : Coreference. (http://www.cogsci.ed.ac.uk/~poesio/MATE/coreference.html)

Hawkins, J.A. (1978). *Definiteness and Indefiniteness*. Atlantic Highlands, NJ: Humanities Press.

Ide N. & Romary, L. (2002). Standards for Language Resources. *Proceedings of the Third Language Resources and Evaluation Conference (LREC)*, Las Palmas, Canary Islands, Spain, 839-44.

Ide N. & Romary L. (2004), International Standard for a Linguistic Annotation Framework. *International Journal on Natural Language Engineering*, forthcoming.

Müller Ch., Strube M. (2001). Annotating Anaphoric and Bridging Relations with MMAX. In: *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, Aalborg, Denmark.

Poesio M., (2000). Coreference. MATE Dialogue Annotation Guidelines-Deliverable 2.1, January 2000, 126-182.(http://www.ims.unistuttgart.de/projekte/mate/

Poesio M., Vieira R., (1998). A Corpus-Based Investigation of Definite Description Use. *Computational Linguistics*, vol. 24, n°2, p 183-216.

Salmon-Alt S. (2001). Du corpus à la théorie : l'annotation (co-)référentielle. *Traitement Automatique des Langues (T.A.L.), n°42/2*, Hermès, Paris.

Tutin A., Trouilleux F., Clouzot C., Gaussier E., Antoniadis G., Zaenen A. (2000). Building a large corpus with anaphoric links in French: some methodological issues. *Actes de Discourse Anaphora and Reference Resolution Colloquium*, Lancaster, UK.

van Deemter K., Kibble R. (2000). On Coreferring: Coreference Annotation in MUC and related schemes, Computational Linguistics 26.4, pp. 615-623

Vieira R., Salmon-Alt S., Gasperin Caroline, Schang E., Othero G. (2002). Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus. *DAARC 2002*, Lisboa, Portugal, September 2002.