Towards a Reference Annotation Framework.
Susanne Salmon-Alt, Laurent Romary

To cite this version:

HAL Id: halshs-00005022
https://halshs.archives-ouvertes.fr/halshs-00005022
Submitted on 22 Nov 2005

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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.


[1] When do we have <ref:de ID="de_01">orange juice</ref:de> at Elmirâ? We have <ref:de ID="de_02">orange juice</ref:de> at Elmirâ at 6 a.m.
<ref:link type="ident" href="coref.xml#id(de_02)"/>
<ref:anchor href="coref.xml#id(de_01)"/>
</ref: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. 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 inflexion 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">verter</feat>
  <feat type="gender">feminine</feat>
  <feat type="number">singular</feat>
</struct>
```

Figure 2: GMT instance

```
<formeFléchie>
  <orth>verter</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.

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 Reference 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 anaphora 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).
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 (informant) and level of confidence (confidence level) of the corresponding information.

3.3.1. Data Categories related to Markables

Besides 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 morphosyntactic 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.). On the other hand, it has been shown for several languages that acceptable inter-annotator agreement could

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 objective and lexical relations. Objectual 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 hypernymy, 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):

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):

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 morphosyntactic representation whereas the integration of multimodal 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


