Data Categories for a Normalized Reference Annotation Scheme.

Susanne Salmon-Alt, Laurent Romary

Data Categories for a Normalized Reference Annotation Scheme

Susanne Salmon-Alt*, Laurent Romary**

* ATILF – CNRS
Nancy, France

** LORIA – INRIA
Vandœuvre-lès-Nancy, France

Susanne.Salmon-Alt@atilf.fr, Laurent.Romary@loria.fr

Abstract

This paper compiles a core set of data categories to be used in combination with the Reference Annotation Framework. It shows also that the underlying objective – ensuring internal and external coherence – is a difficult issue, since reference annotation articulates various levels of linguistic description which all contribute to the identification or qualification of markables and referential links.

1. Current practice of reference annotation

Reference annotation associates referring expressions – usually certain types of noun phrases – with information that enables their interpretation (e.g. antecedents). This type of knowledge is required for a variety of language processing applications, including information extraction and retrieval, machine translation and human-machine dialogue. Reference annotation in a broad sense, covering coreference, anaphora and reference, has already been the subject of substantial practical and theoretical work which suggest basic principles for coherent coding procedures and attempts to unify existing practices (Hirschman & Chinchor 1997, Davies & Poesio 2000, van Deemter & Kibble 2000, Vieira et al. 2003, Salmon-Alt & Romary 2004, Poesio 2004).

In conjunction with the maturity in the field of reference annotation, there is an opportunity to stabilize the corresponding knowledge as an international standard in the context of the recently created ISO committee TC37/SC4 on language resource management. Indeed, this committee aims at providing generic standards for the representation of linguistic data at various levels. It has drafted a set of guidelines for the design of annotation schemes in the domain of linguistic resources (LAF, Linguistic Annotation Framework). Within this framework, we have derived a Reference Annotation Framework (RAF), which is intended to encompass most of the features used in existing practices of reference annotation (Salmon-Alt & Romary 2004).

As a follow-up, this paper briefly recalls the main features of RAF and places the emphasis on a coherent set of linguistic descriptors (or *data categories* in the sense of current ISO TC37 work) intended to fit the needs of a wide variety of reference, coreference and anaphora annotation. In fact, the analysis of existing annotation schemes (for an overview, see Davies & Poesio 2000) shows high discrepancies in the use of specific selections of descriptors. For instance, information attached to the expressions to be annotated varies from morphological to semantic information (Figure 1). The relationships used to type the links between those markables are even more heterogeneous: as illustrated in Figure 2, they include information relative to markables rather than links (*numerical pronoun*, *proper name*), confuse relations between referents and concepts (*part-whole* vs. *conceptual bridging*) and encode relations belonging to other levels of description than reference (*agent*, *role in event*).

type of pronoun (personal, possessive, indefinite etc.), type of noun (common, proper name), determiner (indefinite, definite, demonstrative), syntactic function (subject, object etc), semantic properties (concrete, abstract), localization (number of sentence, paragraph etc)

Figure 1 : Example of current descriptors for markables

identity, coreference, bridging, part-whole, associative, indirect anaphor, unfamiliar, conceptual bridging, set-subset, cause, inferable-of-complement, propositional, possessive, implicit argument, ellipsis, plural NP, numerical pronoun, proper noun, bound anaphor, function-value, instantiation, agent, patient, cause, other-anaphor, role in event…

Figure 2 : Example of current descriptors for relationships

Figure 3 gives an idea about how the descriptors may be combined in a concrete annotation project: here, markables are tagged as <RS> elements and contain attributes for part of speech, gender, number, syntactic functions and syntactic constituents, and semantic type. Links are encoded by an external <LINK> element, with descriptors identifying the arguments of the relation and typing the link. This example points also to some problematic issues: the encoding of complex antecedents (encoded implicitly by order and number of the linked arguments), ambiguity or uncertainty of the human annotator (encoded by question marks), and combining different type of relational information (TYPE for referential or lexical relationships and TYPEANA for the distinction between anaphor and cataphor). Whereas the structural properties of RAF are intended to cover the first two points, the main focus of this paper is the third question, i.e. the coherent use of data categories relevant to reference markables and links.

Incertitude technique, d'abord, tant il est difficile de discerner les implications à moyen et long terme de l'explosion <RS ID="11.1" CAT="GND" MORPGR="FEM" MORPNB="PLU" SYNT1="CNOM" SYNT2="PH1" TYPE="OBJET">***des technologies***</RS> *de l'information et de l' émergence*<RS ID="11.2" CAT="GNI" MORPGR="FEM" MORPNB="SING" SYNT1="CNOM" SYNT2="PH1" TYPE="ABST">***d'une infosphère*** </RS>*dont M. de Saint-Germain souligne - incertitude supplémentaire - qu'*<RS ID="11.3" CAT="PPV" MORPGR="FEM" MORPNB="PLU" SYNT1="C0" SYNT2="PH2" TYPE="IND">***elles***</RS> *sont pilotées par le marché civil.* <LINK ARGS=""???11.1" "11.2" "11.3"" TYPE="COREF" TYPANA="AG"/>

Figure 3 : Example of practice (Clouzot et al. 2000)

2. A meta-model for reference annotation

2.1. The Linguistic Annotation Framework

Our model for specifying and representing reference annotation schemes here is based on the general principles of LAF (Ide & Romary 2004). These principles have already been implemented for the representation of terminological data, in the context of designing an ISO standard (ISO 16642, 2003). Basically, they consider a class of semi-structured documents that can be specified through the combination of a meta-model that informs general practices in organizing information in a given application domain on the one hand, and a selection of data categories (DCS) that characterizes the elementary information units attached to the components of the meta-model, on the other hand. In this perspective, the components in the meta-model are indeed considered as elementary linguistic abstractions which reflect the granularity of the description intended by the meta-model.

2.2. The Reference Annotation Framework

From these general principles it is possible to derive a meta-model that covers the features characterizing reference annotation. Briefly, the reference annotation scheme meta-model (cf. Figure 4), 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 metadata associated with the annotation file, it contains *markables* and *referential links.*

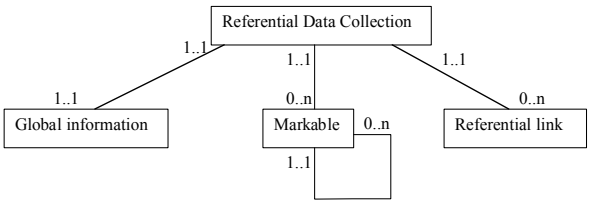

Figure 4: Meta-model for reference annotation.

The basic constituents of any reference annotation scheme are source markables as input, and links to target markables as output. *Markables* are either built upon parsed text chunks (e.g. noun phrases) or directly annotated in the source text. Depending on the underlying theory, they represent anaphora and antecedents (Clouzot et al. 2000), coreferring expressions (Hirschman & Chinchor 1997) or referring expressions and referents (Bruneseaux & Romary 1997). For RAF, markables are the elementary units involved in anaphoric, coreferential, or referential links. They may point to externalized source data (e.g. words, morpho-syntactic units, syntactic chunks), from which relevant linguistic information (type of NP, gender, number) may be percolated. However, they are autonomous and represent essential linguistic abstractions from source data in two senses. First, they are not necessarily isomorphic to elements from the source data. This 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 descriptors specific to the reference level (see section 3).

In addition to markables, any reference annotation schema makes use of (mostly typed) *links* between source and target markables. Those links represent a relation which is necessary for correct discourse interpretation: depending on the theory, this could be an equivalence relation (coreference) or not (reference and anaphora). 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 than just by a pointer attached to the source markable. However, an autonomous link element is preferable for representing ambiguities and different links from the same source markable (Davies & Poesio, 2000). Therefore, RAF introduces a *Referential Link* component, relating markables that are linked by a specific relation.

3. Data Categories for reference annotation

Data categories – as elementary information units attached to the markables and links of the meta-model – play a crucial role in the implementation of any RAF-conform annotation scheme. Not only do they bear the actual information content conveyed by an annotation, but they also characterize the variation of scope for different annotation schemes. Still, in order to make data categories truly usable across applications, several issues must be taken into account.

First, one should provide a stable background for the description and maintenance of data categories in a normative framework. This is the objective of ongoing efforts within ISO TC37 to deploy a data category registry. It ensures that two annotation schemes pointing to the same data category expect reliably the same underlying concept.

Second, a precise structure for describing data categories should exist, so that their semantics and conditions of use can be easily made available to users. A general background is currently provided by ISO 11179 (*Metadata registries*), and ISO TC37 is working on more specific requirements applicable to language resources (current revision of ISO 12620: *Data categories*). For RAF, we use a subset of those requirements to identify, for each data category, an identifier, a definition, a source (when applicable), a profile (to indicate the main linguistic domain that the data category is relevant to) and a conceptual domain (to indicate the list of values, when applicable). These data categories are specifically intended to gather up a first reference list of possible data categories to be used for reference annotation. In the remainder of this section, we address the two issues of (a) markable and (b) referential link related data categories and provide justifications as to their use in a concrete annotation scheme.

Finally, it is essential to provide implementers with precise guidelines that may help them to consistent selections of data categories in the context of a specific annotation project. For example, not all combinations of these categories make sense in the context of reference annotation. Although it is not the main scope of this paper, we will give some hints about the factors to take into account (section 4).

3.1. Markables

This section gives a presentation of some important descriptors associated with reference markables, either genuine to the reference level or to be percolated from lower linguistic description levels and possibly.

3.1.1. Basic descriptors

The /SOURCE TEXT/ data category allows for recovering elementary linguistic data and corresponds to the main mechanism by which an annotation may be related to underlying linguistic data. It should actually be shared by all standards defined under ISO TC37/SC4. As can be seen in the following example, the reference to linguistic data can be expressed in two ways: one can directly embed the source text in the markable (1), thus making the annotation file an autonomous resource, or point to another annotation file by expressing a range in the 'target' attribute (2). In the second case, the source text can be recovered by following the pointer to the primary content, possibly through intermediate annotation levels.

1. <struct id="m_1" type="markable">
   <feat type="source text">la poire</feat>…
   </struct>
2. <struct id="m_1" type="markable">
   <feat type="source text" target="w_2..w_3"/>…
   </struct>

In addition to the identification of a source string, it might be useful to introduce a data category for the concept of minimal matching string such as proposed in MUC (Hirschman & Chinchor 1997), in which a special attribute is used in the answer key to indicate the minimum string that the system under evaluation must include in order to receive full credit for its output.

3.1.2. Referential descriptors

On the very referential description level, relevant information to be associated with markables pertains to characteristics of the referents and is concerned with cardinality, natural gender, definiteness and informational status.

/CARDINALITY/ – denoting the size of a set of referents – and /NATURAL GENDER/ – specifying whether a referent is biologically male or female. It is important to remind that cardinality is distinct from grammatical number, in the sense that it specifies the exact quantify of a noun whereas grammatical number is usually vague. In the same way, natural gender is distinct from grammatical gender by the fact that the latter requires agreement between nouns and modifiers, whereas natural gender does not. Since these rules also apply to anaphoric references, there might be a discrepancy between gender of an antecedent and a corresponding anaphor (3).

3. *Le ministre de la défense* effectue de nombreux déplacements en province. En général, *il*/*elle* est accueillie chaleureusement.

/DEFINITENESS/ whose relations to reference have been extensively discussed by Hawkins (1978) and is a category concerned with the grammaticalization of identifiability and nonidentifiability of referents on the part of the speaker or the addressee. It takes conceptual range values such as /DEFINITE IDENTIFIABILITY/, /INDEFINITENESS/, /GENERIC TERM/, /NON SPECIFIC TERM/ and /SPECIFIC TERM/ and has been used for automatic resolution of definite description by Vieira & Poesio (2000).

The relationship between reference resolution and /INFORMATIONAL STATUS/ – reflecting the speakers assumption about the addressee's knowledge, basically expressed as new/old distinction – goes back to Prince (1981). More recently, Nissim et al. (2004) proposed a first annotation scheme for information status and annotated a corpus using a basic set of data categories /OLD/, /MEDIATED/ and /NEW/, which has been proved to be reliable with respect to inter-annotator agreement.

An additional data category could be /REFERENTIAL STATUS/, with a conceptual range of /PENDING/ or /SOLVED/ to indicate whether a the reference of referring expression has been calculated or not.

3.1.3. Morpho-syntactic and semantic descriptors

- **Part of speech**

/PART OF SPEECH/ – one of the traditional categories of words intended to reflect their functions in a grammatical context – is one of the basic morpho-syntactic information acting as an input to any reference, coreference or anaphora resolution algorithms (Hobbs 1978, Lappin & Leass 1994, Mitkov 1998, Vieira & Poesio 2000). First, it participates in the identification of syntactic chunks as markables on the reference level (filtering out pleonastic pronouns, for example). Second, it encodes crucial grammatical clues generally considered to trigger specific procedures for coreference and anaphora resolution, depending for example on the determiner (indefinite, definite etc), on the type of the head noun (common vs. proper noun) and on subcategorization of pronouns (pleonastic, personal, reflexive etc.).

- **Grammatical gender, number and person**

Similar to part of speech, grammatical gender and number are mandatory data categories to be taken into account for reference annotation and resolution. /GRAMMATICAL GENDER/ distinguishes classes of nouns reflected in the behavior of associated words and takes as conceptual range values such as /MASCULINE/, /FEMININE/, /NEUTER/ and /COMMON/. /GRAMMATICAL NUMBER/ is a category that specifies the quantity of a noun or affects the form of a verb or other parts of speech depending on the quantity of the noun to which it refers. The conceptual range of number encompasses /SINGULAR/, /PLURAL/, /DUAL/, /TRIAL/, /PAUCAL/ and /COLLECTIVE/. Both grammatical gender and number impose basic agreement constraints that are part of any resolution algorithm.

Finally, we mention the necessity to consider /GRAMMATICAL PERSON/, with the conceptual range of /FIRST PERSON/, /SECOND PERSON/ and /THIRD PERSON/ as a major factor for the analysis of deixis in discourse.

- **Syntactic category**

Descriptors related to syntactic aspects are not always present in existing reference annotation schemes. Besides, their actual values vary highly depending on the theoretical background that has led, for instance, to the syntactically annotated resource, upon which the reference

annotation project is built. In this section we thus only do a quick survey of what can be identified as central to reference annotation, without being too precise at this stage with regards their actual definitions and values. /SYNTACTIC CATEGORY/ is usually used in reference annotation to indicate the underlying nature of the source referring expression (in most case a noun phrase), but applies also to the target expression which might be different from a noun phrase (4). For automatic reference resolution, this kind of information can also be useful to estimate accessibility, depending for example on whether the antecedent is embedded in prepositional phrases or not (Leass & Lappin 1994). Although little agreement may be expected as to the values of this data category, one could actually recommend to use neutral values such as /NOUN PHRASE/, /PREPOSITIONAL PHRASE/, /CLAUSE/ etc.

4. Despite the latest negative results, *doctors are still convinced that Tamoxifen can prevent breast cancer*. *This* is because of the way it blocks the action of oestrogen, the female sex hormone that can make the breast cells of some women go out of control.

- **Syntactic function**

/SYNTACTIC FUNCTION/ is a central data category for reference annotation. It is one of the important factors in intra-sentential focus mechanisms and is usually included in reference resolution theories (Reinhart 1983, Grosz & al 1995) and algorithms (Hobbs 1978, Lappin & Leass 1994, Mitkov 1998). Still, the actual values to be considered for this data category may differ highly from an annotation scheme to another. Here again, for the sake of best interoperability between schemes and thus comparability of results, we recommend keeping to basic grammatical relations (/SUBJECT/, /DIRECT OBJECT/, /INDIRECT OBJECT/, /OBLIQUE OBJECT/ etc.), keeping in mind that such a choice may be troublesome in many linguistic situations (e.g. for the qualification of *bunsetsus* in Japanese depending on their case marker).

- **Lexical information**

Lexical information is related to /WORD SENSE/. Word sense disambiguation (WSD) is useful for solving cases of anaphora realized through different head nouns, implying a dictionary look up for synonymy (indirect anaphora) or meronymy (bridging anaphora), as it has been shown by Poesio et al. 2002 or Vieira et al. 2003.

When word senses are not available or not formalized enough to be used in reference resolution, more generic information might characterize a markable. One example is /ABSTRACTNESS/ such as annotated in the introduction (Figure 3) by the 'type' attribute of the <RS> element. Cross-lingual classification experiments on concrete vs. abstract features of the head nouns involved in demonstrative anaphora and the related antecedents have indeed shown that concrete demonstratives have high tendency to take NPs with concrete head noun as antecedents (Vieira et al. 2003) This observation is important for anaphor resolution heuristics, since it allows for excluding less plausible antecedent candidates for concrete demonstratives.

/ANIMACY/ also plays a crucial role for reference resolution. In English, for example, it is a central property for the correct use of personal pronouns. For French, it has been shown that plural pronouns without textual antecedents tend to refer by default to animate referents. Furthermore, animacy is a relevant feature for deriving semantic constraints on reference and anaphora resolution in ambiguous contexts (Ge et al. 1998).

/COLLECTIVENESS/ applies to nouns that stand for a single entity made up of more than one animate creature. Because these nouns behave as both herd animals and solitary creatures, collective nouns can be grammatically either singular or plural, depending on whether the action involving them holds in unison for all members of the group or not. Therefore, coreference to entities introduced by collective nouns may be realized by either plural or singular forms:

5. The *jury* agrees that the state prosecutors did not provide enough evidence, so *its* verdict is not guilty.
6. The *jury* disagree about the guilt of the accused and have told the judge that *they* are hopelessly deadlocked.

In addition to the fact that there exist languages with obligatory possessive markers for the expression of inalienable nouns, linguists consider generally /(IN)ALIENABILITY/ as a relevant feature for referential analysis, related especially to the use of associative anaphora: Kleiber (2001) for example formulates the hypothesis that a referent of a (stereotypical) associative anaphor requires the condition of being an alienable part from the referent of its antecedent. This condition explains the difference in the use of the French definite determiner in example 7, where *église* is an alienable part of *village*.

7. Ils traversait un village. **Son/L'église* était détruite.

/COUNTABILITY/, with a conceptual range of /COUNT NOUN/ and /MASS NOUN/, not only determines whether a noun can became plural and the range of possible determiners associated with it, but explains also – in combination with the alienability issue – the possibility of using well formed associative anaphora (Kleiber 2001). More precisely, the (alienable) part of an antecedent's referent denoted by an associative anaphor has to fulfill the constraint of ontological congruency, i.e. to be of the same countability type as the corresponding whole. This constraint excludes for example associative reference to a material or a property of an antecedent (8):

8. *Pierre* enleva sa casquette. ? *La calvitie* plut à tout le monde.

Finally, /(NAMED) ENTITY CATEGORIZATION/, as a part of semantic information, has also been used for different purposes of reference resolution (Vieira & Poesio 2000, Modjeska 2002).

- **Semantic Roles**

The annotation of /SEMANTIC ROLE/ – the underlying relationship that a participant has with the action referred to by the main verb of the clause – takes an important part in classical referring algorithms (Mitkov 1998). It is argued that a referent being the agent or inter-sentential role parallelism may induce specific biases for the selection of the antecedent. It is also a central issue for studies related to verb-noun anaphora, in case an implicit participant of an event (e.g. *driving*) is made explicit in the following discourse ([agent] *driver*). There is currently a debate as to whether the description of semantic roles should rely on a global classification or be rather centered on (more reliable) local roles, such as those used in a project like FrameNet (Atkins et al. 2003; e.g. /buyer/, /object being bought/ etc. for to *buy*). In fact, the choice of specific semantic roles for the design of a reference

annotation scheme should essentially ensure the consistency with other information sources such as an accompanying semantic annotation scheme or a reference lexicon providing a stable background for the corresponding annotations.

3.2. Links

3.2.1. Basic descriptors

Just in the same way as /SOURCE TEXT/ articulates the markable component of the RAF metamodel with the data to be annotated, there is a need to introduce specific data categories to relate links and markables. This issue is intimately related to the design principles of RAF, where we made the choice of not introducing any a priori dependency between markables and links. We thus use two specific data categories to be anchored on the referential link component. /REFERENTIAL SOURCE/ points to the markable corresponding to the initiating referring expression and as such is mandatory and unique. /REFERENTIAL TARGET/ points to the referring expression being the antecedent (in a broad sense) of the referential source. We exclude more than one /REFERENTIAL TARGET/, since we argue in favor of structural solutions for encoding ambiguity (see Salmon-Alt & Romary 2004 for discussion).

3.2.2. Linguistic and objectal relations

Previous work on reference annotation has shown the need of typing the relation between the linked markables. However, as pointed out by van Deemter & Kibble (2000), reference annotation in the sense considered here (covering coreference and anaphora) must 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, Salmon-Alt & Vieira 2002).

We follow here van Deemter & Kibble (2000) in defining coreference as an equivalence relation expressing identity of referents and anaphora as a relation of interpretational dependency. Table 1 shows that these two relations may or may not co-occur. Current practice covers them both, but does not make a clear distinction. For instance, MUC-7 uses *coreference* for any kind of link. On the other hand, the same combination (non anaphoric coreference, for example) has been characterized differently, depending on whether the framework focus on the relation that holds between the referents of the annotated expressions (*coreference*) or between the referring expressions themselves (*linguistic bridging*, *NP predication*). As a consequence, a consistent annotation framework has to encode the two relations by two different data categories with different conceptual ranges. Our main concern is therefore to distinguish between /OBJECTAL RELATION/ and /LINGUISTIC RELATION/. Objectal relations hold between referents and include /OBJECTAL IDENTITY/, /(IN)ALIENABLE PART OF/ and /SUBSET/ relations. Linguistic relations hold between (parts of) referring expressions (e.g. head nouns) and include /LEXICAL IDENTITY/, /SYNONYMY/, /HYPERNYMY/, /HYPONYMY/ and /MERONYMY/. Table 1 gives an idea about how the simultaneous use of both may help to encode in a more coherent way the previously discussed examples.

However, the precise definition of the conceptual range and the scope of objectal and linguistic relations still needs to be discussed. With respect to the heterogeneity of the data in Figure 2, open questions are in particular the association of some of the information with markables rather than with links (*numerical pronoun*, *other-anaphor*, *agent*); the scope of relationships to be considered (should non anaphoric and non coreferential relations such as *function-value* be covered by a reference annotation framework ?); the proper delimitation of objectal and linguistic relations (to which level pertains *instantiation* or *possession* ?); the definition of a set of constraints on combining values for conjoint objectal and linguistic relations (*objectal identity* excludes *meronymy*); the granularity of values for conceptual range (*linguistic bridging* vs. *partititve possessive*); the question whether to include a specific data category for explicitly encoding the direction of an interpretational dependency (*anaphor* vs. *cataphor*).

+ coreference + anaphor	*a tower…the tower* coreference (MUC-7, Ucrel) identity of reference (Drama) identity (MATE)	R_O : obj. identity R_L : lex. identity
– coreference + anaphor	*the boy…his hair* coreference (MUC-7, Ucrel) linguistic bridging (Drama) partitive possessive (MATE)	R_O : inalien. part R_L : meronymy
+ coreference – anaphor	*Tony Blair… the Prime Minister* identity (MATE) coreference (MUC-7) NP predication (UCREL)	R_O : identity R_L : NP predicat.
– coreference – anaphor	*the temperature… 90 degrees* coreference (MUC-7) function-value (MATE)	R_O : funct. value R_L : ?

Table 1 : Objectal and linguistic relationships

4. Towards the design of coherent schemes

The description of possible data categories to be used in a RAF-conform annotation scheme shows that the task of compiling a specific subset may not be as straightforward as it could seem when considering simple annotation tasks limited to, for instance, coreference chains. More precisely, there are two sets of relevant factors for designing an annotation scheme, related to the respectively external and internal coherence of the data category selection.

First, the actual choice of data categories to be attached to the RAF metamodel is related to the underlying annotated data upon which reference annotation is anchored. Depending on whether this data has been morpho-syntactically annotated or is limited to plain text, descriptors should not be necessarily duplicated from one level to another. Furthermore, the actual usage of the resulting annotation is also important. For instance, the case another layer of named entity annotation is available, one would avoid overloading the reference

annotation scheme with too many concept related data categories. Besides, we consider that more work should be done on the study of inheritance/percolation mechanisms across annotation scheme to make sure that lower level information is seamlessly accessible to higher levels (Ide et al. 2000). As a whole, the design of a reference annotation scheme should be considered in the wider context of the specific corpus and/or application in order to limit the data category selection to a meaningful subset without interference with other levels and to ensure easy understandability for annotators.

The second issue – internal coherence – is actually more of a conceptual problem than it is technical. This is related to the fact that reference annotation is in essence a place of articulation between linguistic information and higher level semantic/pragmatic aspects. From the point of view of data categories, this appears immediately when both /GRAMMATICAL GENDER/ and /NATURAL GENDER/ occurs in the same annotation scheme. It is also the case when the complementarity of two sets of data categories such as /FIRST PERSON/ + /SECOND PERSON/ and /SPEAKER/ + /ADDRESSEE/ correspond to the actual interpretation of the linguistic marker in context. Finally, it becomes even more complex when semantics has to be taken into account: for instance, a low-level data category such as /GRAMMATICAL NUMBER/ should be used coherently by semantic features related to /COUNTABILITY/ and /COLLECTIVENESS/ (whether these are linguistically marked or not) and articulated with purely objectal information such as /CARDINALITY/. The same problems occurs for combining linguistic and objectal relationships for referential link.

As an overall conclusion, this paper is a first attempt in compiling a core set of data categories that may be used in combination with the RAF metamodel to design specific reference annotation schemes. The underlying objective – ensuring the best level of interoperability between them – is a difficult issue since reference annotation articulates various levels of linguistic description that all contribute to the identification or qualification of markables and referential links. One of the trade-offs is therefore to make sure that data categories pertaining to other domains than reference properly speaking can be defined independently of there actual use in RAF conformant annotation schemes while providing precise guidelines for their use in reference annotation tasks.

The next steps in the attempt to standardize reference annotation are (a) stabilization of the main data categories identified in this paper as contributions to the ISO TC37 DCR by working in collaboration with the joint ISO-ACL Sigsem working group on semantic content representation; (b) providing precise mappings of existing reference annotation schemes to those data categories and identifying possible discrepancies or missing descriptors; and (c) submitting the RAF metamodel and general principles as a preliminary draft for a future international standard within ISO TC37/SC4.

5. References

Atkins S., Fillmore C., Johnson C. (2003). Lexicographic Relevance: Selecting Information From Corpus Evidence. *International Journal of Lexicography*, 16(3): 251-280.

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

Hirschman L., Chinchor N. (1997). MUC-7 coreference task definition. *MUC-7 Proceedings*.

Clouzot C., Antoniadis G., Tutin A. (2000). Toward Automatic Generation of Understandable Pronouns in French Language. *Lectures Notes in Artificial Intelligence,* 1835:242-252.

Ge N., Hale J., Charniak E. (1998). A statistical approach to anaphora resolution. *Proc. of the 6th Workshop on Very Large Corpora*.

Grosz B., Joshi A. and Weinstein S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics,* 12(2) :203-225.

Hawkins J. (1978). *Definiteness and Indefiniteness*. Croom Helm, London.

Hobbs J. (1978). Resolving Pronoun References. *Lingua*, 44:311-338.

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

Kleiber G. (2001). *L'anaphore associative*. PUF, Paris.

Lappin S., Leass H. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535-561.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. *Proc. COLING 98/ACL'98*.

Modjeska N. (2002). Lexical and Grammatical Role Constraints in Resolving *Other*-anaphora. *Proc. of DAARC 2002.*

Nissim, M., Dingare, S., Carletta, J., and Steedman, M. (2004). An Annotation Scheme for Information Status in Dialogue. *Proc. of LREC 2004*, Lisbon.

Davies S., Poesio M. (2000). Coreference. *MATE Dialogue Annotation Guidelines-Deliverable D2.1,* January 2000, 126-182.

Poesio M. (2004). The MATE/GNOME Proposals for Anaphoric Annotation, Revisited. *Proc. of the 5th SIGdial Workshop*.

Poesio M., Ishikawa T., Walde S., Vieira, R. (2002). Acquiring lexical knowledge for anaphora resolution. *Proc. of LREC 2002*.

Poesio M., Vieira R. (1998). A corpus-based investigation of Definite Description Use. *Computational Linguistics*, 24(2):183-216.

Prince E. (1981). Toward a taxonomy of given-new information. In Cole P., *Radical Pragmatics*, Academic Press, New York, 223-255.

Reinhart T. (1983). Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy* 6(1):47-88.

Salmon-Alt S., Romary L. (2004). RAF - towards a reference annotation framework. *Proc. of LREC 2004*.

van Deemter K., Kibble R. (2000). On Coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics* 26(4).

Vieira R., Salmon-Alt S., Gasperin Caroline, Schang E., Othero G. (2003). Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus. In : *Anaphora Processing*. A. Branco et al. (eds.), John Benjamins Publishing Company.

Vieira R., Poesio M. (2000). An Empirically-Based System for Processing Definite Descriptions, *Computational Linguistics*, 26(4):525-579.