INTERNATIONAL STANDARD

ISO 24617-9

First edition
2019-12

Language resource management — Semantic annotation framework —

Part 9: Reference annotation framework (RAF)

Reference number
ISO 24617-9:2019(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2. www.iso.org/directives

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received. www.iso.org/patents

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management*.

A list of all parts in the ISO 24617 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

This document is intended to complement the ISO 24617 series and to provide all the necessary conceptual and technical mechanisms for the annotation of referential phenomena in multimodal discourse. Reference phenomena are an essential component for the understanding and structuring of discursive mechanisms, ranging from very basic pronominal relation to complex bridging anaphora. Annotating such phenomena in an interoperable way improves the re-usability of language resources in such applications in language technology as named entity recognition, text understanding and synthesis, text summarization, information retrieval, automatic question-answering, man-machine dialogue, and machine translation.

The content of this document builds upon various projects and software platforms that have been dealing with reference annotation (RA), in particular the following References [9],[2],[16],[21], [26],[25],[22],[5],[15],[13] but also the TEI P5 guidelines. Based on these and other previous works, the Referential Annotation Framework (RAF) aims at providing a synthesized way of treating various reference phenomena in discourse. In continuity with most practices in the field, RAF focuses on marking up referring expressions in a discourse and the relations that hold between them and the corresponding entities, whether this is based upon employing crowd sourcing or machine learning strategies.

Language resource management — Semantic annotation framework —

Part 9: Reference annotation framework (RAF)

1 Scope

This document provides a comprehensive model for the annotation and representation of referential phenomena in natural language texts and multimodal interactions. Such phenomena can cover simple anaphoric or coreferential mechanisms as well as more complex bridging or multimodal mechanisms. It provides a reference serialisation in XML defined as a customisation of the TEI P5 guidelines. In addition, the document describes the core data categories related to referential entities and link structures, and also needed for the description of annotation schemes and serialisation mechanisms for implementing conformant models as concrete data formats.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 24622-1, *Language resource management — Component Metadata Infrastructure (CMDI) — Part 1: The Component Metadata Model*

TEI P5, *Guidelines for Electronic Text Encoding and Interchange*. Version 3.5.0. Last updated on 29th January 2019. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Extensible Markup Language (XML) 1.0 (Fifth Edition), W3C Recommendation 26 November 2008. https://www.w3.org/TR/REC-xml/

IETF BCP 47, *Tags for Identifying Languages*, September 2009. https://tools.ietf.org/html/bcp47

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1
anaphora
linguistic mechanism by which the interpretation of a *referring expression* (3.7) depends on another expression mentioned in the same text or discourse

Note 1 to entry: The notion of anaphora is more general than that of *coreference* (3.3): the interpretation of anaphora is context-dependent, whereas coreference is determined rather rigidly independently to its possible use of context (see Reference [25]).

Note 2 to entry: The term is used in this document in its general sense since, for instance, no specific distinction is made here with the notion of cataphora (i.e. coreference) with a more specific expression occurring later in a discourse).

3.2
communicative segment
elementary portion of a multimodal interaction

3.3
coreference
identity of *referents* (3.6) of two referring expressions

Note 1 to entry: The concept covered here corresponds to the data category *objectal identity,* described in Annex A.

3.4
objectal relation
relation between two *discourse entities* (3.6) reflecting their intended association from a referential point of view

Note 1 to entry: The referential association may identify that they are identical, disjoint, or overlapping, or that one includes the other (see References [6] and [25]).

3.5
reference
relation between a referring expression and a *discourse entity* (3.6) denoted by it

Note 1 to entry: The verb "to refer to" expresses such a relation: if there is a reference relation between an expression x and a discourse entity *e*, then x is said to refer to *e*.

3.6
referent
discourse entity

extra-linguistic entity which is denoted, or pointed out, by a *communicative segment* (3.2)

Note 1 to entry: discourse entity is used preferably in the context of the description of the concrete syntax whereas referent is used in the abstract syntax, but also when the underlying process is implied by the expression.

3.7
referring expression
communicative segment (3.2) that specifically designates an entity or an event, whether concrete or abstract, discourse new or old, real or fictional

4 Basic principles

This document provides a generic framework for the annotation of reference phenomena in discourse, whether in textual, spoken or multimodal form. As required by ISO 24612 and ISO 24617-6 principles, its syntax is formulated at two levels, abstract and concrete. The abstract syntax characterizes in abstract terms what RAF theoretically is. There can be a variety of concrete syntaxes that conform to a proposed abstract syntax. XML-serialization is the most commonly accepted concrete syntax among them.

The proposed serialisation is entirely conceived as a customisation of the TEI P5 guidelines and builds upon the existing constructs provided by ISO 24611 for morpho-syntactic annotation. Any implementation of the present document shall also be compliant with the TEI P5 guidelines and consequently the XML W3C recommendation.

As suggested by [25], this document focuses on the annotation of referring expressions such as noun phrases in a language as its markable expressions, abbreviated as "markables". This includes entities (John, the dog) as well as events, as expressed through noun phrases (the party, the meeting). Verbal expressions denoting events may be marked as well, however, since they also may refer to events. For example, “We met, and it lasted all morning.” It leaves out annotation of non-referring noun phrases and

bound anaphora involving quantification to some extent. It does not address such tasks as annotation of the relation between a subject and a predicative noun phrase (e.g., "**John** is **a singer and guitar player**"). Nor does it treat type coreference. This includes so-called sloppy identities (e.g., "John loves his wife and **so does** Bill") and verb-phrase anaphors (e.g., "Animals **suffer** as much as we **do**", "Peter **cuts** vegetables much faster than I **do** (cut vegetables)") in general. In delimiting its markables, RAF attempts to make clear the theory of reference as much as possible without getting into theoretical details and also the notion of coreference against a more general notion of anaphora.

5 Meta-model for reference annotation

5.1 Overview

The general meta-model for reference annotation is presented in Figure 1. It articulates the identification and qualification on two complementary levels:

— the linguistic level where *referring expressions* can be segmented and qualified within the flow of a discourse;

— the discourse domain where *discourse entities* referred to by referring expression are identified as relevant for modelling the discourse domain.

Both objects may be further refined by data categories and links among them as described further on in this document.

Referring expressions are also anchored on *communicative segments*, which may be linguistic segments as well as any multimodal communicative sign (gesture, face movement, etc.) that is relevant for the identification of the referring act.

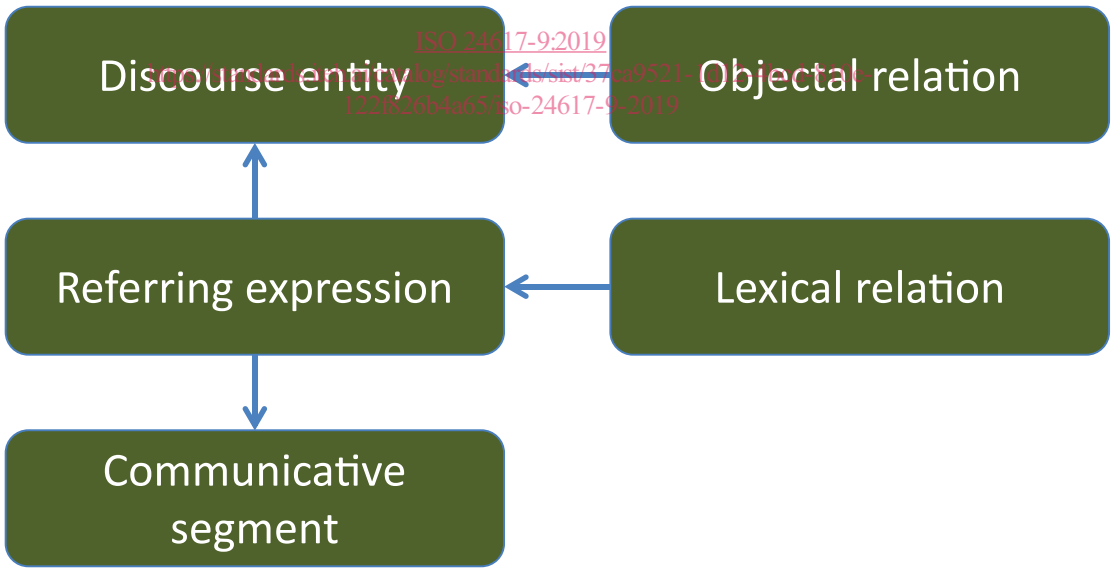

Figure 1 — Meta model for reference annotation

5.2 Referring expressions

The referring expression component corresponds to the identification of one or several communicative segments in the textual source as well as within other multimodal channels (visual or auditory) that can be interpreted as a single referring act. A referring expression may for instance correspond to a single continuous linguistic segment.

EXAMPLE 1 [en] I ate [**the apple**]$_i$.

where the referring expression *i* is a single definite description.

It can also be the combination of simpler referring expressions as is the case within a coordination.

EXAMPLE 2 [en] I ate [[**an apple**]$_i$ **and** [**an orange**]$_j$]$_k$,

where the referring expressions *i* and *j* are part of the larger referring expression *k*.

It can also be expressed by one or several sub-token markers, as is the case in agglutinative languages or when referring morphemes are bound within another token.

EXAMPLE 3 [it] prendo[lo]$_i$ (I take it.).

Depending on the serialisation, referring expressions can be represented as explicitly recursive, by means of links among them, or implicitly recursive, by systematically pointing to their occurrences in the source text.

Markables for reference annotation, however, include complex anaphors, zero pronouns, and discourse deixis. Plural pronouns such as "they" may have partial antecedents, as illustrated by Example 4 below, while zero pronouns often occur in conversations in some languages other than English, as illustrated by a Korean example below in Example 5. Discourse deixis such as "this" and "that" refer to part of what has been said in discourse. Spatial and temporal deixis such as "here", "there", "now", and "then" are also to be marked up as referring expressions.

EXAMPLE 4 [en] **John**$_i$ married **Lisa**$_j$ yesterday and **they**$_{\{i,j\}}$ went to Paris for **their**$_{\{i,j\}}$ honeymoon.

EXAMPLE 5 Dialogue in Korean [ko]: "**Mia** wass-ni?" (Did Mia come?)

"Yey, wass-e-yo". (Yes, [**pro**] came.)

NOTE The subject in the answer is implied and represented in the translation as a zero pronoun [pro].

EXAMPLE 6 [en] I don't believe that **this story of his** is true.

Markables are not restricted to referring expressions of nominal and pronominal forms. They may also cover verbal (anaphoric) forms such as "so do(es)" or "do", as in the following examples.

EXAMPLE 7 [en] Mary loves her husband and **so does** Jane.

EXAMPLE 8 [en] Animals **suffer** as much as we **do**.

5.3 Data categories for referring expressions

Referring expressions may be characterised by a variety of data categories that are felt to be relevant for the annotation project at hand. These categories may percolate from lower annotation levels (e.g. morpho-syntactic, syntactic or semantic) or specifically relate to the occurrence context of the referring expression. The following data categories may be considered as the basis for the characterisation of referring expressions. When the corresponding data category is not defined in another ISO standard, the definitions provided in Annex A shall be adopted.

— Morpho-syntactic categories relevant for referring expressions resulting from the percolation of one or several properties of the components of the referring expression: grammatical gender (*grammaticalGender*, ISO 24611), grammatical number (*grammaticalNumber*, ISO 24611), person (*person*, ISO 24611).

— Syntactic or semantic data categories resulting from the identification and qualification of the referring expression as a syntactic constituent: syntactic category (*syntacticCategory*, ISO 24615-1[1)]),

1) With typical values such as *nounPhrase* and *verbPhrase* (ISO 24615-1).

grammatical case (*grammaticalCase*, ISO 24611), grammatical function (*grammaticalFunction*, ISO 24615-1).

— Semantic-pragmatic data categories: referential status, definiteness (*definiteness*, ISO 24611), animacy.

EXAMPLE 9 [en] $\text{Lee}_{\text{feminine,i}}$ loves [$\text{her}_{\text{feminine,i}}$ husband]$_{\text{masculine,j}}$ but $\text{he}_{\text{masculine,j}}$ doesn't care.

5.4 Lexical relations

Lexical relations can be associated with data categories expressing lexical semantic relations that usually form the basis of the referential interpretation process. These data categories define relations between lexical items or, by inheritance from their nominal heads, nominal phrases. For reference annotation, the relations that are defined between lexical items can be extended to larger linguistic units, such as noun phrases. The data categories provided in Annex A cover the most commonly needed cases: synonymy, hyponymy, hypernymy, compatibility, meronymy, and lexical identity.

EXAMPLE 10 [en] John bought **a pear**_i and Jane **an apple**_j, for they love **these fruits**_{i,j}. [hyponymy, together with a subset relation at discourse entity level].

5.5 Discourse entities

The data categories associated with discourse entities concern properties of extra-linguistic entities involved in the interpretation of referring expressions. These properties are marked grammatically in some languages, for example animacy and alienability. The core properties elicited in this document are the following ones:

— *abstractness*: A complex data category which can take two values: *abstract* and *concrete;*

— *alienability:* A complex data category which can take two values: *alienable* and *inalienable;*

— *animacy:* A complex data category which can take two values: *animate* and *inanimate;*

— *cardinality:* the provision of the number of entities within a discourse entity interpreted as a set.

— *entity categorisation:* A complex data category that allows the linking of a discourse entity to an underlying classification or ontology

— *natural gender:* the provision of the natural gender for a discourse entity seen as a living entity; precise definitions and sources are available in Annex A.

5.6 Objectal relations

Objectal relations are relations between discourse entities seen as extra-linguistic concepts. The following relations[25],[23],[24] form the basis of the present standard in this respect:

— *objectal identity*, to express an exact coreference relation;

— *part of,* when a discourse entity is identified as being a component of another one;

— *member of,* when a discourse entity is identified as an element within a set of referents;

— *subset*, when a discourse entity is seen as a set of entities all part of a larger set.

Precise definitions and sources are available in Annex A.

5.7 Metadata

The metadata for reference annotation documents contains global information concerning annotator(s), tool, date, and pointer to scheme specification such as DCS (Data Category Selection). It can also

include local information concerning inter-annotator agreement, confidence level with respect to tools, revisions, and updates.

For the specification of such metadata, implementation shall comply to the TEI P5 guidelines or ISO 24622-1. It may also comply to the OLAC (Open Language Archive Community) initiative.

6 Abstract syntax, concrete syntax, and semantics of annotations

6.1 Introduction

In this document, referential annotations are defined in accordance with the principles of semantic annotation laid down in ISO 24617-6. Accordingly, annotations have a three-part definition consisting of an abstract syntax, a concrete syntax, and a semantics. The abstract syntax defines *annotations* in the sense of the Linguistic Annotation Framework (ISO 24612), namely as a specification of linguistic information that is added to segments of source data, independent of the format in which the information is represented. For semantic annotation, such specifications are pairs, triples and in general n-tuples of semantic concepts. ISO 24612 defines *representations,* by contrast, as the rendering of annotations in a particular format. A concrete syntax specifies a representation format for the annotation structures defined by the corresponding abstract syntax. Finally, a semantics is defined for the annotations defined by the abstract syntax, allowing alternative representation formats to share the same semantics.

The present clause specifies first the abstract syntax of reference annotations, subsequently their semantics, and finally a concrete syntax for representing annotations as a customisation of the TEI P5 guidelines. The TEI P5 guidelines provide a generic XML vocabulary for the representation of textual content and associated annotations. In representing various relevant features of referring expressions, discourse entities and the relations between them, this document follows ISO 24610-1, as required by ISO 24612.

6.2 Abstract syntax

The structures defined by an abstract syntax are n-tuples consisting of basic concepts, taken from a store of such concepts called the 'conceptual inventory', or (nested) n-tuples of such structures. Two types of structure are distinguished: *entity structures* and *link structures*. An entity structure contains semantic information about a segment of primary data; link structures contain information about the way two or more such segments are semantically related.

6.2.1 Conceptual inventory

The conceptual inventory of RAF is a 6-tuple: <M, RF, GP, RStat, ORels, LRels>, where

1. M is a set of markables;
2. RF is a set of referential features of discourse entities;
3. GP is a set of grammatical properties of referring expressions;
4. RStat ('referential status') is a pragmatic property of discourse entities;
5. ORels is a set of objectal relations;
6. LRels is a set of lexical relations.

In line with the metamodel shown in Figure 1, the abstract syntax distinguishes two kinds of entity structure, viz. for discourse entities (objects and events) and for referring expressions, and two kinds of link structure, one for relating discourse entities and one for relating referring expressions.

