SLOVENSKI STANDARD
oSIST ISO/DIS 24617-9:2019

01-oktober-2019

Upravljanje jezikovnih virov - Ogrodje za semantično označevanje - 9. del: Referenčni okvir označevanja (RAF)

Language resource management -- Semantic annotation framework -- Part 9: Reference annotation framework (RAF)

Gestion des ressources linguistiques -- Cadre d'annotation sémantique -- Partie 9: Référence (ISOref)

Ta slovenski standard je istoveten z: **ISO/DIS 24617-9:2019**

ICS:

01.020	Terminologija (načela in koordinacija)	Terminology (principles and coordination)
35.240.30	Uporabniške rešitve IT v informatiki, dokumentiranju in založništvu	IT applications in information, documentation and publishing

oSIST ISO/DIS 24617-9:2019 **en,fr,de**

DRAFT INTERNATIONAL STANDARD

ISO/DIS 24617-9

ISO/TC **37**/SC **4**

Secretariat: **KATS**

Voting begins on:
2019-02-05

Voting terminates on:
2019-04-30

Language resource management — Semantic annotation framework —

Part 9:
Reference annotation framework (RAF)

Gestion des ressources linguistiques — Cadre d'annotation sémantique —

ICS: 01.020

This document is circulated as received from the committee secretariat.

Reference number
ISO/DIS 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

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 on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT) see the following URL: Foreword - Supplementary information

ISO 24617-9 was prepared by Technical Committee ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management, WG 2 Semantic annotation*.

ISO 24617 consists of the following parts, under the general title *Language resource management — Semantic annotation framework*:

Part 1: Time and events (SemAF-Time, TimeML))

Part 2: Dialogue acts (SemAF-DA)

Part 3: Named entities (SemAF-NE)

Part 4: Semantic roles (SemAF-SR)

Part 5: Discourse structures (SemAF-DS)

Part 6: Principles of semantic annotation (SemAF Principles)

Part 7: Spatial information

Part 8: Semantic discourse relations (SemAF-SDR)

Part 9: Reference Annotation Framework (RAF)

Introduction

This document is intended to complement the ISO 24617 series (Language resource management -- Semantic annotation framework (SemAF)) and 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 will improve 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: Hirschman & Chinchor (1997)'s MUC-7 Coreference Task Definition (CDT), Bruneseaux & Romary (1997), Poesio et al. (1999)'s MATE meta-scheme, Poesio & Davies (2000), Poesio & Vieira (2000), van Deemter & Kibble (2000), Salmon-Alt (2001), Müller & Strube (2001), Vieira et al. (2002), Byron & Gegg-Harrison (2004), Poesio (2004)'s GNOME, Passoneau (1996)'s DRAMA, Müller & Strube (2006)'s MMAX2-based annotation scheme, Pustejovsky et al. (2013)'s Brandeis annotation scheme, but also the TEI guidelines (*TEI P5*). 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.

As suggested by van Deemter & Kibble (2000), RAF 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 (NPs) 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 NP (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.

This document also has benefited from the in depth work carried out within the EU project e-Content Lirics e-Content/Lirics (2004 – 2006; http://lirics.loria.fr/).

Language resource management – Semantic annotation framework – Part 9: Reference Annotation Framework (RAF)

1 Scope

This document aims at providing a comprehensive model for the annotation and representation of referential phenomena in natural language texts and multimodal interactions. Such phenomena may 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 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, in whole or in part, are normatively referenced in this document and are indispensable for its application. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 24610-1:2006, *Language resource management — Feature structure — Part 1: Feature structure representation (FSR)*

ISO 24612:2012, *Language resource management — Linguistic annotation framework (LAF)*

ISO DIS 24617-6:2015, *Language resource management — Semantic annotation framework — Part 1: Principles of semantic annotation (SemAF-Principles)*

ISO 24611:2012 Language resource management — Morpho-syntactic annotation framework (MAF)

TEI Consortium, eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. [Version number]. [Last modified date]. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ ([Date of access]). ⇐ Note: to be completed when finalising the standard

The Unicode Standard (6.0 ed.). Mountain View, California, USA: The Unicode Consortium. ISBN 978-1-936213-01-6. http://www.unicode.org/versions/Unicode6.0.0/

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

3 Terms and definitions

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

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

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

Note: terms corresponding to data categories are not mentioned here, see annex A for a full documentation of the normative data categories introduced by this document.

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* (3.2): the interpretation of anaphora is context-dependent, whereas *coreference* is determined rather rigidly independently to its possible use of context (see van Deemter & Keeble (2000)).

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* (3.2) with a more specific expression occurring later in a discourse).

3.2

communicative segment
elementary portion of a multimodal interaction

3.3

coreference
equality of **referents** (3.6) of two linguistic 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 Cruse, 1986 and van Deemter and Kibble, 2000)

3.5

reference
relation between a linguistic 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

3.7

referring expression
Communicative segment that specifically designates an entity or an event, whether concrete or abstract, discourse new or old, real or fictional

5 Basic requirements

RAF provides a generic framework for the annotation of reference phenomena in discourse, whether in textual, spoken or multimodal form. As required by ISO 24612 LAF and ISO 24617-6 SemAF-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 guidelines and builds upon the existing constructs provided by ISO 24611 for morpho-syntactic annotation.

6 Meta-model for reference annotation

6.1 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.

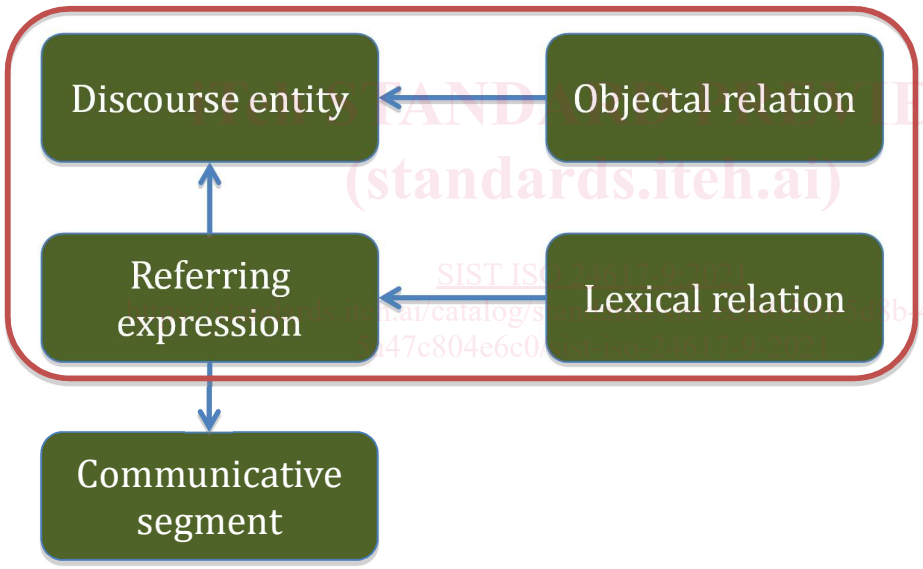

Figure 1: Meta model for reference annotation

6.2 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, e.g.:

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, e.g.:

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*.

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

ISO DIS 24617-9:2018(E)

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 3 below, while zero pronouns often occur in conversations in some languages other than English, as illustrated by a Korean example below in example 4. 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 3: [en] **John**$_i$ married **Lisa**$_j$ yesterday and **they**$_{\{i,j\}}$ went to Paris for **their**$_{\{i,j\}}$ honeymoon.

Example 4: 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 5: [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 example:

Example 6: [en] Mary loves her husband and **so does** Jane.

Example 7: [en] Animals **suffer** as much as we **do**.

6.3 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, its normative definition is provided in Annex A:

- 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]), grammatical case (*grammaticalCase*, ISO 24611), grammatical function (*grammaticalFunction*, ISO 24615-1);
- Semantic-pragmatic data categories: referential status, definiteness (*definiteness*, ISO 24611), animacy

Example 8: [en] **Lee**$_{feminine,i}$ loves [**her**$_{feminine,i}$ **husband**]$_{masculine,j}$, but **he**$_{masculine,j}$ doesn't care.

6.4 Lexical relations

Lexical relations may 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 may be extended to larger linguistic units, such as noun

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