
**Language resource management —
Semantic annotation framework —**

Part 9:
**Reference annotation framework
(RAF)**

STANDARDSISO.COM : Click to view the full PDF of ISO 24617-9:2019



STANDARDSISO.COM : Click to view the full PDF of ISO 24617-9:2019



COPYRIGHT PROTECTED DOCUMENT

© ISO 2019

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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	iv
Introduction	v
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 Basic principles	2
5 Meta-model for reference annotation	3
5.1 Overview	3
5.2 Referring expressions	3
5.3 Data categories for referring expressions	4
5.4 Lexical relations	5
5.5 Discourse entities	5
5.6 Objectal relations	5
5.7 Metadata	5
6 Abstract syntax, concrete syntax, and semantics of annotations	6
6.1 Introduction	6
6.2 Abstract syntax	6
6.2.1 Conceptual inventory	6
6.2.2 Annotation structures: Entity structures and link structures	7
6.3 Semantics	8
6.3.1 Discourse entity structures and objectal relation links	8
6.3.2 Referential expression entity structures and lexical relation links	9
6.4 Implementing an XML serialisation compliant with the TEI P5 guidelines	10
6.4.1 Introduction	10
6.4.2 Namespace	10
6.4.3 Generic principles attached to a TEI compliant serialisation	10
6.4.4 Feature structures	11
6.4.5 General document architecture	12
6.5 Implementation of the <i>Referring expression</i> component	12
6.6 Implementation of the <i>Discourse entity</i> component	13
6.7 Implementation of referential relations	13
6.8 Objectal relations: grouping	14
6.9 Alternative linking: ambiguity	15
6.10 Multiple links	15
6.11 Representing referential chains	16
6.12 Bridging phenomena	16
Annex A (normative) Data categories for reference annotation	18
Annex B (informative) Complementary examples or partial examples referred to in the main text of the document	25
Bibliography	26

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.

STANDARDSISO.COM : Click to view the full PDF of ISO 24617-9:2019

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.

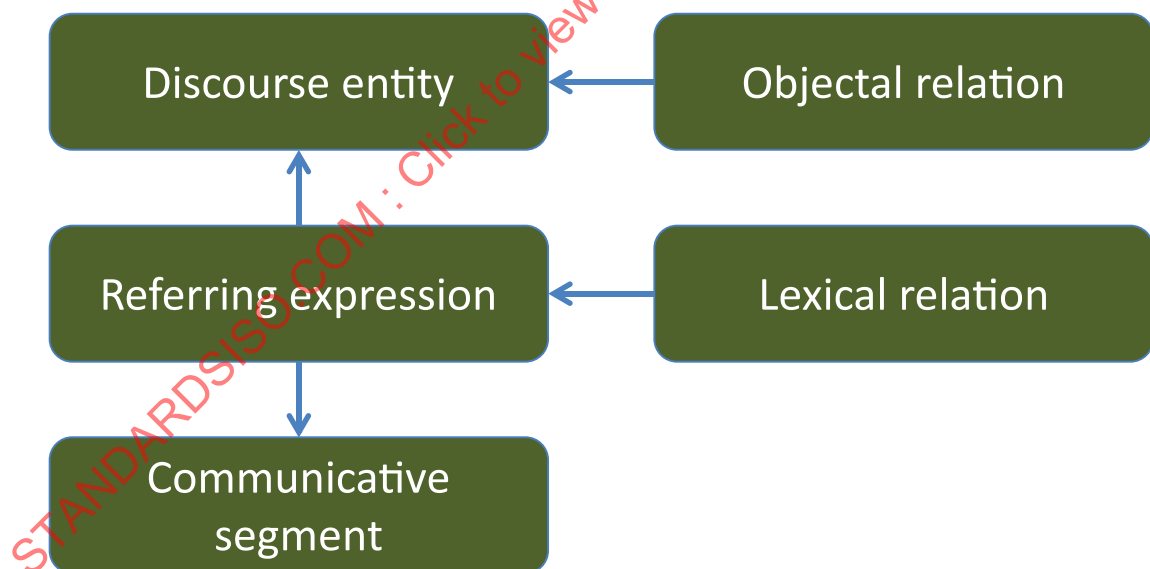


Figure 1 — Meta model for reference annotation

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 24615-1¹⁾),

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] Lee_{feminine,i} loves [her_{feminine,i} husband]_{masculine,j}, but he_{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.

6.2.2 Annotation structures: Entity structures and link structures

Since an entity structure specifies certain semantic information about a segment of primary data, it is formally a pair $\langle m, \sigma \rangle$, consisting of a markable 'm' that identifies the data segment, and the semantic information designated by ' σ '.

In an entity structure for discourse entities, markables are typically noun phrases (possibly inflected or comprising affixes depending on the language family), including complex anaphors, zero pronouns, and discourse deixis, but also expressions that refer to events, or, more precisely, to 'eventualities' (i.e. events, states, or processes, and possibly facts), as illustrated in 5.2. A full-fledged noun phrase consists of two parts: (1) a noun, called the 'lexical head' of the noun phrase, complemented by one or several syntactic dependencies such as adjectives, prepositional phrases or other modifiers, and (2) one or more determiners such as "the", "some", "most", and "less than 2000". The head noun denotes a set of entities, called the *source domain* of the noun phrase.

The noun phrases that are the markables of reference annotation for other than event-type entities are also the markables for the annotation of quantification²⁾. In the abstract syntax for quantification annotation, entity structures consist of (1) the specification of a domain that the quantification is restricted to; (2) the 'involvement', i.e. the specification of how many entities or how much of the domain are/is participating in an event (as in "More than five thousand students protested"); (3) the specification of 'definiteness' (definite or indefinite); and (4) the 'size', i.e. a specification of the size of the domain (as in "Two of the twenty students failed the exam"). These features are also relevant for reference annotation. An entity structure for (non-event) discourse entities is thus a nested quadruple $\langle m, \langle D, q, d, N \rangle \rangle$ (D = domain, q = involvement, d = definiteness, N = domain size).

The domain of a quantification is mostly not the entire noun phrase source domain but rather a contextually determined subdomain formed by certain salient members of the source domain (the ones recently mentioned in the discourse, for example), called the *reference domain*^[22]; its specification includes a stipulation of whether it is a set, a single individual, or a mass quantity (such as a bit of fresh air, some coffee, or some music). The component D in an entity structure is therefore a pair $D = \langle P, i \rangle$ consisting of a predicate that characterizes the domain and the specification of its individuation (i = set, individual, or mass). The predicate that is characteristic for a source domain typically has certain properties that can be specified in an ontology or a domain model, as well as properties that are linguistically determined. Examples of the latter kind are animacy (being animate or inanimate), and natural gender (male, female or additional relevant value for the study at hand), which are grammatically marked in many languages; an example of the former kind is incompatibility (see definition in Annex A). In order for reference annotation not to be dependent on particular ontologies or other external knowledge sources, it can be useful to mark up such properties as additional features. For use in reference annotation, the entity structures for discourse entities as defined for the annotation of quantification are extended with optional semantic features, which include animacy, humanness, alienability, and abstractness. The choice and use of these features depends on the linguistic properties of the language of discourse and of the availability of ontological knowledge sources.

Of special interest for reference annotation are pronominal noun phrases, i.e. noun phrases of which the lexical head is a pronoun, such as "it", "her", "one of them", "some of it", "both of them"; such noun phrases differ from full-fledged noun phrases mainly in that they do not specify a source domain from which the discourse entities are taken. Personal pronouns do however carry some domain-constraining information, which differs from language to language. In English, for example, the pronoun "it" can only refer to a non-human discourse entity, and "he" only to a male human. In the entity structure for the discourse entity of an occurrence of "he", the semantic information is thus $\langle \langle \text{male, individual} \rangle, 1, \text{definite}, 1 \rangle$. The various personal pronouns all have in common that their referents must be salient in the context of use. Semantic features like animacy and humanness are thus of particular interest for pronominal noun phrases.

Link structures for reference annotation are triples $\langle e1, e2, R \rangle$ consisting of two entity structures and a relation between them; either $e1$ and $e2$ are entity structures for discourse entities and R is a (possibly complex) objectal relation, or $e1$ and $e2$ are entity structures for referring expressions and R is a lexical relation.

2) An ISO standard for the annotation of quantification is in preparation, see Reference [3].

An entity structure for a referring expression specifies a lexical head and a number of properties that can be useful for detecting reference relations. These properties are mostly grammatical in nature, such as the syntactic category of the expression, grammatical gender, grammatical number, grammatical case, and grammatical function (subject, object, indirect object). Such an entity structure is thus a nested n-tuple $\langle m, \langle H, C, G, s \rangle \rangle$ (H = lexical head, C = syntactic category, G is set of grammatical properties, s = referential status).

Annotation structures describe the referential structure of a discourse, as defined by the objectal relations between discourse entities and the lexical relations between (the nominal heads of) referring expressions. An annotation structure is formally a set of entity structures and a link structures that connect some or all of the entity structures. A (sub-)set of discourse entity structures of which one member is directly or indirectly linked to all the other members defines a co-reference chain.

6.3 Semantics

6.3.1 Discourse entity structures and objectal relation links

From a semantic point of view, the focus of reference annotation is on the characterization of the objectal and lexical relations among discourse entities and referring expressions, respectively. The semantic characterization of the participants in these relations is within the scope of reference annotation in so far as it is relevant for establishing these relations.

The semantic interpretation of an objectal link structure is a formal relation between two discourse entities. Such relations can be basic objectal relations, viz. *objectal identity*, *part of*, *member of*, *subset of*, *distinct from*, or they can be complex relations that involve these basic objectal relations (examples below) or that constitute cases of metonymy.

The standard semantic interpretation of noun phrases, referring to discourse entities other than events, is that they express ‘generalized quantifiers’, i.e. properties of sets of individuals that participate in eventualities. This interpretation can be captured by a second-order DRS (Discourse Representation Structure, Reference [10]). For example, the noun phrase “Six of the boys” in the sentence “Six of the boys played tennis” corresponds to the DRS $[X \vdash x \text{ in } X \rightarrow \text{boy}_0(x), \text{card}(X)=6]$, where ‘ boy_0 ’ represents the predicate ‘boy’ restricted to those boys that are salient in the given context (as indicated by the use of the definite article “the”).³⁾ The variable ‘ X ’ in this DRS (called a ‘discourse referent’ in Discourse Representation Theory) can be thought of as designating that subset of the reference domain that consists of those boys that were actually involved in playing tennis – this set X is what is also called the ‘referent’ here. The example sentence might the initial part of a discourse that goes as follows:

EXAMPLE 11 “Six of the boys played tennis. When it started to rain and the girls arrived, the boys interrupted their game. Two of them went home.”

The noun phrase “the boys” corresponds to the DRS $[Y \vdash y \text{ in } Y \leftrightarrow \text{boy}_0(y)]$; the occurrence of the NP “Six of the boys” has created a new context where the boys in the set X of six boys are most salient, and therefore the set Y which is the referent of “the boys” is precisely that set of boys: $Y = X$, a case of objectal identity. Similarly, the noun phrase “Two of them” introduces a referent Z , related to X by the objectal relation subset-of.

A referring expression is an expression that has a referent. A simple way of providing a semantics of referential relations is to introduce a function ‘Ref’ that assigns the intended referent to a referring expression, and to make the objectal referential relations in a discourse explicit as relations between the values of the ‘Ref’ function applied to the referring expressions. In Example 11 the referring expressions of interest are m_1 = “Six of the boys”, m_2 = “the boys”, m_3 = “Two of them”. Using the notation m' as short for $\text{Ref}(m)$, the objectal relations can be captured by the triples $\langle m_1', m_2', \text{Objectal Identity} \rangle$ and $\langle m_3', m_1', \text{Subset-of} \rangle$. An example showing more complex objectal relations is the following:

EXAMPLE 12 “Take two apples. Remove their skin. Cut one in slices, mash the other.”

3) The restriction to contextually salient boys can be captured formally by means of a dynamic predicate ‘salient’, and construing boy_0 as $\lambda x. \text{boy}(x) \ \& \ \text{salient}(x)$.

Referring expressions: m1 = “two apples”, m2 = “their skin”, m3 = “one”, m4 = “the other”.

Objectal relations:

1. For all x, member-of(x, m2') --> Exists y, member-of (y,m1') such that part-of(y,x)
2. member-of(m3', m1)
3. member-of(m4', m1')
4. distinct(m4', m3')

The following example shows how complex objectal relations can be, and also that these complex relations can be domain-specific and do not necessarily involve any of the basic objectal relations:

EXAMPLE 13 Press the grapes very gently. Store their juice in the fridge for an hour. Then take half of it and sieve it twice. Put it back into the fridge. Now press the grapes again and collect their juice.

Referring expressions: m1 = “the grapes”, m2 = “their juice”, m3 = “half of it”, m4 = “it”, m5 = “it”, m6 = “the grapes”, m7 = “their juice”.

Objectal relations:

1. first-press(m2', m1')
2. half-of(m3', m1')
3. identity(m4', m3')
4. twice-sieved(m5', m3')
5. identity(m6', m1')
6. second-press(m7', m1')

6.3.2 Referential expression entity structures and lexical relation links

Referential expressions are linguistic objects rather than semantic objects; lexical relations such as *synonymy*, *hyponymy*, *hypernymy*, and *antonymy* denote a relation between the meanings of the lexical items that form the head nouns of two noun phrases. Lexical relations thus apply only to referring expressions that are full-fledged noun phrases (in contrast with pronominal noun phrases).

For two full-noun-phrase referential expressions corresponding to the entity structures E1 = <m1, <H1, C1, G1, s>> and E2 = <m2, <H2, C2, G2, s2>> a lexical relation, such as synonymy, is just a relation between the lexical heads H1 and H2: Synonymy(H1, H2), similarly for the other lexical relations.

Note that on this approach the grammatical features of referring expressions do not have a semantic interpretation (and neither does their pragmatic ‘referential status’), so from a semantic point of view, only the specification of a lexical head is obligatory; all other elements are optional. This is in perfect accordance with the principles of semantic annotation in ISO 24617-6, in which three types of optional elements are distinguished (see Reference [4]): (1) elements that are semantically without impact; (2) elements that can be omitted from an annotation representation because they have a default value in the encoded annotation structure; (3) those that make annotations more informative if present. The grammatical features of referring expressions are all cases of optional elements of type (1).⁴⁾

The rest of this clause first presents generic constraints related to the serialisation of annotation structures, followed by a systematic provision of XML-TEI constructs for the various components of the meta-model.

4) The specification of a reference domain size in a discourse entity structure is a case of an optional element of type (3).

6.4 Implementing an XML serialisation compliant with the TEI P5 guidelines

6.4.1 Introduction

All concrete digital representations shall comply to the general constraints of the XML W3C recommendation as well as to the specific technical requirements from the TEI P5 guidelines. When available, specific links to these are provided in relation to the elicitation of how specific TEI elements are to be used in the context of the present document.

6.4.2 Namespace

In all XML examples in this section, to simplify the actual representations, it is assumed, unless otherwise stated, that XML elements belong to the TEI namespace, with the implicit following namespace declaration at the root element of the corresponding XML document:

```
xmlns=http://www.tei-c.org/ns/1.0
```

6.4.3 Generic principles attached to a TEI compliant serialisation

The TEI P5 guidelines provide several generic mechanisms that either result from it being an XML application or originate from the TEI underlying architecture.

Character encoding is dealt with, as for any XML application, within the XML declaration of the corresponding document. For instance, an XML document starting with the following declaration:

```
<?xml version="1.0" encoding="utf-8"?>
```

indicates that the character encoding used is utf-8, as defined by the Unicode standard^[39].

The *identification of elements* within an XML document shall be made by means of the @xml:id attribute as defined in the XML W3C recommendation.

Pointing to an XML element is made on the basis of the pointing mechanisms defined in chapter “Linking, Segmentation, and Alignment”⁵⁾ of the TEI P5 guidelines, in conformance with ISO 24612. In the rest of this document, examples rely on simple URIs pointing explicitly to XML elements by means of their @xml:id attribute. More complex pointing schemes, involving in particular character offsets are also possible.

The TEI P5 guidelines have taken up the @xml:lang attribute for indicating the *working language* of any content within an XML document. The value of @xml:lang shall be compliant with IETF BCP 47 wherever this information is needed in a document. IETF BCP 47, which is based upon the IANA Language Subtag Registry⁶⁾, integrates the codes from the ISO 639 series for languages and ISO 15924 for scripts as needed.

The annotation mechanisms described in this document operate on a *source linguistic content*, which can either correspond to an encoded text or the transcription of a spoken (even multimodal) interaction. The source content can also be encoded according to the TEI P5 guidelines, with various levels of possible inline descriptive objects. In the case of transcribed spoken content, the source document shall be compliant with ISO 24624. Most examples in this document rely on a simple representation illustrated in Example 14, where the following TEI elements have been used:

- <s>⁷⁾ to indicate a sentence level segmentation;
- <w>⁸⁾ to indicate a token level unit;

5) <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html>

6) <https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry>

7) <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-s.html>

8) <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-w.html>

- `<pc>`⁹⁾ for punctuation marks.

EXAMPLE 14 TEI Encoding of the French sentence “Prendre une poire et la couper” ([en] *Take a pear and cut it*):

```
<s>
  <w xml:id="w0">Prendre</w>
  <w xml:id="w1">une</w>
  <w xml:id="w2">poire</w>
  <w xml:id="w3">et</w>
  <w xml:id="w4">la</w>
  <w xml:id="w5">couper</w>
  <pc>.</pc>
</s>
```

In actual linguistic annotation projects, the reference annotation layer described in this document can point to different types of lower level objects, from a simple sequence of characters up to higher level linguistic units such as morpho-syntactic word forms (word forms in the sense of ISO 24611) or phrases when the corpus has been annotated for syntax (as described in the ISO 24615 series). From a technical point of view, the reference annotation shall rely on the same URI based mechanisms as the one described here.

The representation of referential annotations in TEI are mainly based on the `<span>`, `<interp>`, `<link>` and `<fs>` elements as specified further in this document. These elements can in turn be grouped together within additional elements:

- `<spanGrp>`¹⁰⁾ and `<linkGrp>`¹¹⁾ to group `<span>` and `<link>` elements respectively. When these elements have a `@type` attribute, it is inherited by their child elements (unless superseded);
- `<annotationBlock>`, as described in ISO 24624 (*Transcription of Spoken Language*), when annotations correspond to the same linguistic segment (for instance an utterance or a sentence);
- `<listAnnotation>`, when an annotation scheme groups together annotations of the same type or corresponding to the same workflow, mainly for curation purposes.

6.4.4 Feature structures

In compliance with ISO 24610-1, this document represents the data categories attached to the various components of the meta-model by means of feature structures as follows:

- `<fs>` (*feature structure*) groups together various features attached to the implementation of a given component;
- `<f>` (*feature*) expresses a single feature with its name (in the `@name` attribute) and value (as content of the element);
- Feature values are encoded as a `<symbol>` element whose `@value` attribute indicates the appropriate code.

Feature names are expressed on the basis of the camel case code provided in [Annex A](#) for complex data categories (e.g. *animacy*). Feature values should be expressed on the basis of the camel case code provided in [Annex A](#) for simple data categories (e.g. *animate*). For example:

```
<f name="animacy">
  <symbol value="animate"/>
</f>
```

9) <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-pc.html>

10) <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-spanGrp.html>

11) <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-linkGrp.html>

6.4.5 General document architecture

As illustrated in Example 15 below, the TEI vocabulary allows one to embed standoff annotation as a specific child of the root element of the document to be annotated. The annotations themselves are grouped together within the `<standOff>` element, which contains its own metadata within the `<teiHeader>` element. The various components of the reference annotation model are in turn organized as subsets marked up with `<listAnnotation>` elements, complemented with a `@type` attribute.

Note that Sub-elements can be further structured in `<annotationBlock>` elements when they correspond to the same textual segment (e.g. sentence, utterance, etc.). This is meaningful for referring expressions, whereas the other annotation objects are more likely to correspond to inter-sentential mechanisms.

EXAMPLE 15 General structure of an annotated document:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>...</teiHeader>
  <standOff>
    <teiHeader>...</teiHeader>
    <listAnnotation type="referringExpressions">...<annotationBlock>
      <span type="referringExpression"/>
    </annotationBlock>...</listAnnotation>
    <listAnnotation type="discourseEntities">
      ...<interp type="discourseEntity"/>...</listAnnotation>
    <listAnnotation type="lexicalRelations">
      ...<link type="lexicalRelation"/>...</listAnnotation>
    <listAnnotation type="objectalRelations">
      ...<link type="objectalRelation"/>...</listAnnotation>
    </standOff>
  <text>
    <body>
      <!-- Main document goes here -->
    </body>
  </text>
</TEI>
```

6.5 Implementation of the *Referring expression* component

The Referring expression component is implemented by means of the XML element `<span>`¹²⁾ complemented by the following attributes:

- `@type`, set to the value *referringExpression*, to indicate that the element corresponds to a referring expression;
- `@target`, to point to one or several linguistic segments from the source text. In the case when one wants to point to an interval between two portions of the source text, the attributes `@from`, for indicating the beginning of the interval, and `@to`, for indicating the end of the interval, can replace the `@target` attribute;
- `@ana`, to point to an ISO 24610-1 compliant feature structure implementing the data categories associated with the referring expression.

Example 16 provides an encoding example based on Example 13¹³⁾ where two referring expressions “une poire” (*a pear*) and “la” (*it [feminine]*) have been marked-up. Each referring expression is attached to a feature structure eliciting specific constraints related to it. This illustrates in particular the possible combination of reference-specific data categories, with morpho-syntactic or syntactic data categories, which may have been percolated from lower annotation levels and which can be useful, e.g. for training machine learning processes.

NOTE In the example the value for the `@type` attributes on `<span>` elements are inherited from the `@type` attribute on `<spanGrp>`.

12) <http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-span.html>

13) For the sake of readability, the corresponding linguistic segments have been kept as comments within the `<span>` elements

EXAMPLE 16 Encoding of referring expressions for Example 13:

```

<spanGrp type="referringExpression">
  <span xml:id="m1" from="#w1" to="#w2" ana="#fs1"><!-- une poire --></span>
  <span xml:id="m2" from="#w4" ana="#fs2"><!-- la --></span>
</spanGrp>
<fs xml:id="fs1">
  <f name="syntacticCategory"><symbol value="nounPhrase"/></f>
  <f name="determinerType"><symbol value="indefinite"/></f>
  <f name="referentialStatus"><symbol value="discourseNew"/></f>
</fs>
<fs xml:id="fs2">
  <f name="syntacticCategory"><symbol value="pronominalPhrase"/></f>
  <f name="referentialStatus"><symbol value="discourseOld"/></f>
</fs>

```

6.6 Implementation of the *Discourse entity* component

The Discourse Entity component is implemented by means of the XML element `<interp>`¹⁴⁾ complemented by the following attributes:

- `@type`, set to the value *discourseEntity*, to indicate that the element corresponds to a discourse entity;
- `@inst`, to point to one or several linguistic referring expressions which are considered to refer to this discourse entity;
- `@ana`, to point to an ISO 24610-1 compliant feature structure implementing the data categories associated with the discourse entity.

As further discussed in 6.7 (*referential chains*), it is possible to create a single discourse entity associated with all coreferring expressions as illustrated below (Example 17).

EXAMPLE 17 Example of the representation of a discourse entity associated with two referring expressions “el hombre” (the man) and “sus” (his) from the source example provided in Annex B.2:

```

<listAnnotation type="referringExpressions">
  <annotationBlock target="#u-12-d1e24564">
    <span xml:id="m1" target="#w2 #w3" type="referringExpression">
      <!-- el hombre --></span>
    <span xml:id="m2" target="#w8" type="referringExpression">
      <!-- sus --></span>
    </annotationBlock>
  </listAnnotation>
<listAnnotation type="discourseEntities">
  <interp xml:id="e1" inst="#m1 #m2" type="discourseEntity">
    <!-- HOMBRE --></interp>
  </listAnnotation>

```

6.7 Implementation of referential relations

The Referential links component is implemented by means of the XML element `<link>`¹⁵⁾ complemented by the following attributes:

- `@type`, set to one of the two values *lexicalRelation* or *objectalRelation*, to indicate the nature of the relation expressed either between two referring expressions or two discourse entities;
- `@target`, to point to two referring expressions in the case of lexical relations or discourse entities for objectal relations;

14) <http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-interp.html>

15) <http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-link.html>

- @ana, to point to an ISO 24610-1 compliant feature structure implementing the data categories associated with the corresponding link.

Example 18 below illustrates the implementation of a simple objectal identity relation between two discourse entities, corresponding to the source example provided in [Clause B.2](#).

EXAMPLE 18 Annotation variant from Example 16 with two distinct discourse entities and an objectal relation between them:

```
<listAnnotation type="referringExpressions">
  <annotationBlock target="#u-12-d1e24564">
    <span xml:id="m1" target="#w2 #w3" type="referringExpression">
      <!-- el hombre --></span>
    <span xml:id="m2" target="#w8" type="referringExpression">
      <!-- sus --></span>
    </annotationBlock>
  </listAnnotation>
<listAnnotation type="discourseEntities">
  <interp xml:id="e1" inst="#m1" type="discourseEntity"/>
  <interp xml:id="e2" inst="#m2" type="discourseEntity"/>
</listAnnotation>
<listAnnotation type="objectalRelations">
  <link xml:id="link1" ana="#fs1" target="#e2 #e1" type="objectalRelation"/>
  <fs xml:id="fs1">
    <f name="objectalRelation"><symbol value="objectalIdentity"/></f>
  </fs>
</listAnnotation>
```

6.8 Objectal relations: grouping

Anaphoric mechanisms can comprise the grouping of already mentioned discourse entities as one single set, by means of various mechanisms: plural pronouns, hypernymic definite descriptions, etc. This should be taken into account by means of the *memberOf* objectal relation as illustrated in Example 19 below where a plural pronoun "les" (*them*) can be partially linked to more than one antecedent, for instance, "une poire" (*a pear*) and "une pomme" (*an apple*).

EXAMPLE 19

- (a) Source: [fr] Prendre **une poire** et la couper. Enlever **la peau**. Laver **une pomme**. Éplucher **le fruit**. **Les** faire cuire. Servir **l'une** et **l'autre** avec de la glace.

Translation [en]: take a pear and cut it in two. Peel it. Wash an apple. Peel the fruit. Cook them. Serve the one and the other with some ice cream.

- (b) Annotation of referring expressions

```
<spanGrp type="referringExpression">
  <span xml:id="m1" from="#w1" to="#w2" ana="#fs1">
    <!-- une poire --></span>
  <span xml:id="m4" from="#w10" to="#w11" ana="#fs2">
    <!-- une pomme --></span>
  <span xml:id="m6" from="#w15" to="#w15" ana="#fs3">
    <!-- les --></span>
</spanGrp>
<fs xml:id="fs1">
  <f name="syntacticCategory"><symbol value="nounPhrase"/></f>
  <f name="determinerType"><symbol value="indefinite"/></f>
  <f name="grammaticalNumber"><symbol value="singular"/></f>
<f name="referentialStatus"><symbol value="discourseNew"/></f>
</fs>
<fs xml:id="fs2">
  <f name="syntacticCategory"><symbol value="nounPhrase"/></f>
  <f name="determinerType"><symbol value="indefinite"/></f>
  <f name="grammaticalNumber"><symbol value="singular"/></f>
<f name="referentialStatus"><symbol value="discourseNew"/></f>
</fs>
<fs xml:id="fs3">
```

```

    <f name="syntacticCategory"><symbol value="pronominalPhrase"/></f>
    <f name="determinerType"><symbol value="definite"/></f>
    <f name="grammaticalNumber"><symbol value="plural"/></f>
    <f name="referentialStatus"><symbol value="discourseOld"/></f>
  </fs>

```

(c) Annotation of discourse entities

```

<interp xml:id="e1" inst="#m1" type="discourseEntity"/>
<interp xml:id="e4" inst="#m4" type="discourseEntity"/>
<interp xml:id="e6" inst="#m6" type="discourseEntity"/>

```

(d) Annotation of referential links

```

<linkGrp>
  <link xml:id="link1" type="objectalRelation" ana="#fs4" target="#e1 #e6"/>
  <link xml:id="link2" type="objectalRelation" ana="#fs4" target="#e4 #e6"/>
</linkGrp>
<fs xml:id="fs4">
  <f name="objectalRelation"><symbol value="memberOf"/></f>
</fs>

```

6.9 Alternative linking: ambiguity

Where there are competing interpretations in the way referential expression can relate to preceding ones, it is possible to introduce weighted alternative links by means of the <alt> element¹⁶⁾.

Example 20 below illustrates a case of ambiguity in linking the noun phrase "le fruit" (*the fruit*) which may refer to either "une poire" (*a pear*) or "une pomme" (*an apple*).

EXAMPLE 20

- (a) Source: [fr] Prendre une poire et la couper. Enlever la peau. Laver une pomme. Éplucher le fruit.

Translation [en]: *Take a pear and cut it. Peel it. Wash an apple. Peel the fruit.*

- (b) Annotation: referring expressions, discourse entities and alternations between objectal relations

```

<spanGrp type="referringExpression">
  <span xml:id="m1" from="#w1" to="#w2" ana="..."><!-- une poire --></span>
  <span xml:id="m4" from="#w10" to="#w11" ana="..."><!-- une pomme --></span>
  <span xml:id="m5" from="#w13" to="#w14" ana="..."><!-- le fruit --></span>
</spanGrp>
<interpGrp from="#w1" to="#w2" ana="..." type="discourseEntity"/>
  <interp xml:id="e1" inst="#m1" ana="..." />
  <interp xml:id="e4" inst="#m4" ana="..." />
  <interp xml:id="e5" inst="#m5" ana="..." />
</interpGrp>
<linkGrp>
  <link xml:id="link1" type="objectalRelation" ana="#fs3" target="#e5 #e1"/>
  <link xml:id="link2" type="objectalRelation" ana="#fs3" target="#e5 #e4"/>
</linkGrp>
<alt mode="excl" target="#link1 #link2" weights="0.5 0.5"/>
<fs xml:id="fs3">
  <f name="objectalRelation"><symbol value="objectalIdentity"/></f>
</fs>

```

6.10 Multiple links

In many situations, the interpretation of a referring expression relies on the existence of a complex context from which the actual referent may be extracted. When one wants to elicit such phenomena (see Reference [22]), it becomes necessary to rely on more than one link associated with a given discourse unit.

16) <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-alt.html>

Example 21 shows how a referential markable such as "l'autre" (*the other*) has several links with different objectal relations. It may be linked to "une pomme" (*an apple*), "les" (*them*), or "l'une" (*the one*) each with a different objectal relation, namely *coreference*, *subset*, or *referential disjunction*, respectively.

EXAMPLE 21

- (a) Source: Prendre une poire et la pocher dans du vin rouge et des épices. Faire cuire une pomme au four avec du sucre et de la cannelle. Laissez-les refroidir légèrement. Servir l'une et l'autre avec de la glace à la vanille.

Translation: Take a pear and cook it in some red wine with spice. Cook an apple with sugar and cinnamon in the oven. Let them warm down a little. Serve the one (fruit) and the other with vanilla ice cream.

- (b) Annotation: Link group with different objectal relations

```
<linkGrp type="referentialLink">
  <!-- l'autre => une pomme -->
  <link xml:id="link1" ana="#fs1" target="#e8 #e4"/>
  <!-- l'autre => les -->
  <link xml:id="link2" ana="#fs2" target="#e8 #e6"/>
  <!-- l'autre => l'une -->
  <link xml:id="link3" ana="#fs3" target="#e8 #E7"/>
</linkGrp>
<fs xml:id="fs1">
  <f name="objectalRelation"><symbol value="objectalIdentity"/></f>
</fs>
<fs xml:id="fs2">
  <f name="objectalRelation"><symbol value="subset"/></f>
</fs>
<fs xml:id="fs3">
  <f name="objectalRelation"><symbol value="referentialDisjunction"/></f>
</fs>
```

6.11 Representing referential chains

Referential chains correspond to sequences of referring expressions, which an annotator, or an annotation process, interprets as all referring to the same discourse entity. Depending on the complexity of the corresponding annotation purpose, this document provides three methods for representing referential chains, with different associated expressive powers:

- the baseline for marking up referential chains is to have a discourse entity point to all the referring expressions considered as referring to the same entity;
- when several types of objectal relations are used within a given annotation scheme, a discourse entity shall be explicitly associated with each referring expression and objectal identity links shall be elicited between these. For instance:

EXAMPLE 22 Simple coreferential link between two markables (on the basis of Examples 14 and 16).

```
<link xml:id="link1" type="objectalRelation" ana="#fs1" target="#m2 #m1"/>
<fs xml:id="fs1">
  <f name="objectalRelationType"><symbol value="objectalIdentity"/></f>
</fs>
```

6.12 Bridging phenomena

The annotation of bridging anaphora, i.e. an anaphoric link between two referring expressions that reflects a non-direct lexical (e.g. hyponymy, see Reference [21]) or objectal (e.g. element-set, part-whole, subset, substance, causation) association can vary quite extensively from one annotated corpus to another^[17].

This document does not provide a specific definition of the complex notion related to bridging anaphora, but provides all the necessary mechanism to make it possible for specific applications to implement

their intended implementation of the phenomena. In essence, bridging anaphora can be optimally implemented by combining the appropriate data categories for lexical and objectal relations on the basis of the normative ones described in [Annex A](#), namely:

- for lexical relations: synonymy, hyponymy, hypernymy, meronymy;
- for objectal relations: objectal identity, part of, subset, memberOf.

The categories for *substance* and *causality* have not reached enough consensus and are left to implementers to define.

STANDARDSISO.COM : Click to view the full PDF of ISO 24617-9:2019

Annex A (normative)

Data categories for reference annotation

A.1 General

This annex contains a structured list of categories that shall be used when annotating the corresponding mechanism within a reference annotation scheme. It describes, for each category, its name, a camel case representation, a conceptual domain (for complex data categories), a definition (with its source(s)), one or several examples (with their source(s) when applicable), as well as some usage notes if needed.

The camel case ("*camelCase*") form shall be used in any implementation of this document. In particular, it is the basis for all values of the @type attribute or for names and values of features in the TEI based serialization described in [6.4](#).

A.2 Properties of referring expressions

Beyond the morpho-syntactic and syntactic data categories that can be percolated from lower levels of annotations (e.g. morphosyntactic and syntactic), the data categories described in this section shall be used to qualify referring expressions.

A.2.1 Referential status

Camel case form: referentialStatus

Conceptual domain: {/discourseNew/, /discourseOld/}

Definition: qualifies the status of the referring expression with regards to contextual discourse

Source: References [\[18\]](#) and [\[19\]](#).

A.2.2 Discourse old

Camel case form: discourseOld

Definition: refers to entities which have already been mentioned in the current discourse

Source: Reference [\[18\]](#).

A.2.3 Discourse new

Camel case form: discourseNew

Definition: refers to entities which have not already been mentioned in the current discourse

Source: Reference [\[18\]](#).

A.3 Lexical relations

A.3.1 Linguistic referential relation

Camel case code: linguisticReferentialRelation

Conceptual domain: {/sameHead/, /pronominal/, /synonymy/, /hyponymy/, /hypernymy/, /compatibility/, /antonymy/, /incompatibility/, /acronymy/, /meronymy/, /metonymy/}

Definition: Pattern of association between linguistic units X and Y in a language, established on truth-conditional relations between parallel sentences in which X and Y occupy identical structural positions.

Source: Adapted from References [30] and [6].

Note: In case of reference annotation, lexical relations can be annotated between the heads and modifiers of noun phrases.

A.3.2 Same head relation

Camel case code: sameHead

Definition: Relation between two lexical units X and Y, which are syntactically identical (i.e. two noun phrases for instance) and share the same lexical head.

Source: Adapted from [27] and [14].

Examples: [en] {an iron flashlight, the flashlight}, [en] {the violin, this new violin}, [de] {zwei Hunde, diese Hunde}, [br] {tad-kozh, tad-kozh a-berzh mamm}

[en] She bought a *violin* / She plays with his *new violin* very well.

[gcf] lè *vyann la* ja kuit[...]/pa tèlman bizwen asèzoné vrèman *vyann la* ('when the meat is cooked/no need to heavily season the meat'), Adapted from Schang, Antoine & Lefevre-Halftermeyer (2017).

A.3.3 Pronominal relation

Camel case code: pronominal

Definition: Relation between two lexical units X and Y, where Y is a pronoun.

Source: Adapted from [27] and [14], from [6].

Examples: [de] {ein Freund, er}, [en] {a flashlight, it}, [en] {the fiddle, it}, [br] {tad-kozh, hemañ}

[en] She bought a *new violin* to play Irish folk music / She plays *it* very well.

[gcf] I bizwen fè *on kolombo*/ki jan yo ka fè *sa*? ('she must make a *colombo* [dish]/How can you do *this*?), Adapted from Schang, E., Antoine, J. Y., & Lefevre-Halftermeyer, A. (2017). Les chaînes coréférentielles en créole de la Guadeloupe. In *TALN'2017, atelier DILITAL*.

A.3.4 Incompatibility

Camel case code: incompatibility

Definition: Exclusion relation between classes. When X and Y have the same superordinate, but they are disjunct classes, X and Y are said to be incompatible.

Source: Adapted from [7].

Examples: [*A cat*]_i jumped into the house. [*The dog*]_j started to bark.

Note: this notion can also be called *co-hyponymy*.

A.3.5 Compatibility

Camel case code: compatibility

Definition: Relation between classes. When X and Y have the same superordinate, and their denotations overlap.

Source: Adapted from [7].

A.3.6 Synonymy

Camel case code: synonymy

Definition: Relation between two lexical units X and Y which are syntactically identical and have the property that any declarative sentence S containing X has equivalent truth conditions to another sentence S' which is identical to S, except that X is replaced by Y

Note: Synonymy is the linguistic parallel of the *identity* relation between classes. Synonyms differ in peripheral traits, related for example to stylistic, dialectal or diachronic variations.

Source: Adapted from [6].

Examples: [de] {Hund, Köter}, [en] {flashlight, torch}, [en] {glad, joyful, happy}, [en] {violin, fiddle}

[en] He plays the *violin* very well/He plays the *fiddle* very well.

A.3.7 Hyponymy

Camel case code: hyponymy

Definition: Relation between lexical units X and Y characterised by the property that the sentence *This is a(n) X* entails, but is not entailed by the sentence *This is a(n) Y*.

Note: Hyponymy and its converse hypernymy are the linguistic parallels of the relation of inclusion between two classes.

Source: Adapted from [6].

Examples: [en] *animal/dog*, *red/scarlet*, *to kill/to murder*

[en] This is a *dog*. => This is an *animal*.

[en] This is a *scarlet* flower. => This is a *red* flower.

A.3.8 Hypernymy

Camel case code: hypernymy

Definition: Relation between lexical heads X and Y characterised by the property that the sentence *This is a(n) Y* entails, but is not entailed by the sentence *This is a(n) X*.

Source: Adapted from [6].

Example: *dog/animal* (*animal* is a hypernym of *dog*)

Alternate term: hyperonymy

Note: Hyperonymy is the converse of hyponymy.

A.3.9 Meronymy

Camel case code: meronymy

Definition: An inclusion relation between lexical heads X and Y which reflect a potential part-whole relation between their referents in discourse.

Source: Adapted from [7], p. 140.

Example: *finger:hand* (finger is said to be a meronym of hand, and hand is said to be the holonym of finger).

A.3.10 Metonymy

Camel case code: metonymy

Definition: Y as the subpart or related part of an object X "stands for" the object itself.

Source: Adapted from [20]

Example: *L'omelette au jambon est partie sans payer*. See Reference [11]

A.4 Properties of discourse entities

A.4.1 Abstractness

Camel case code: abstractness

Conceptual domain: {/abstract/, /concrete/}

Definition: Attribute of a semantic entity stating whether or not it is considered as a physical object or consisting of matter.

Source: Adapted from Reference [30]

A.4.2 Abstract

Camel case code: abstract

Definition: Property of a semantic entity, stating that it is not considered as a physical object or consisting of matter.

Source: Adapted from [30]

Example: *la chance, freedom*

A.4.3 Concrete

Camel case code: concrete

Definition: Property of a semantic entity, stating that it is considered as a physical object or consisting of matter.

Source: Adapted from [30]

Example: *a farmer, bread, water*

A.4.4 Animacy

Camel case code: animacy

Conceptual domain: {/animate/, /inanimate/}