INTERNATIONAL STANDARD

ISO 24615-1

First edition
2014-02-01

Language resource management — Syntactic annotation framework (SynAF) —

Part 1: Syntactic model

Gestion de ressources langagières — Cadre d'annotation syntaxique (SynAF) —
Partie 1: Modèle syntaxique

Reference number
ISO 24615-1:2014(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web 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 (see 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 (see 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.

The committee responsible for this document is ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management*.

This first edition of ISO 24615-1 cancels and replaces ISO 24615:2010, of which it constitutes a minor revision.

ISO 24615 (all parts) is designed to coordinate closely with ISO 24612, *Language resource management — Linguistic annotation framework (LAF)*, ISO 24613:2008, *Language resource management — Lexical markup framework (LMF)*, and ISO 24611, *Language resource management — Morpho-syntactic annotation framework*.

ISO 24615 consists of the following parts, under the general title *Language resource management — Syntactic annotation framework (SynAF)*:

— *Part 1: Syntactic model*

The following part is under preparation:

— *Part 2: XML serialization (<Tiger2/>)*

Introduction

ISO 24615 is based on numerous projects and pre-standardisation activities that have taken place in the last few years (see Abeillé, 2001[9]), to provide reference models and formats for the representation of syntactic information, whether as the output of a syntactic parser, or as annotations of language resources (treebanks). For several years, the Penn Treebank initiative has served as a *de facto* standard for treebanking, but more recent works e.g. the Negra/Tiger initiative (see: http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/) in Germany or the ISST initiative in Italy [see Montemagni (2003) [18]] demonstrate the viability of a more coherent framework that can account for both (hierarchical) constituency and dependency phenomena in syntactic annotation.

The eContent project "LIRICS", has been seminal in gathering a group of experts, who initiated the ISO 24615 (SynAF) project. While preparing SynAF, this group confirmed that existing initiatives indeed share a common data model that offers a good basis for the SynAF metamodel (see the study made in Deliverable D.3.1 "Evaluation of initiatives for morpho-syntactic and syntactic annotation" of the EU project LIRICS, available at http://lirics.loria.fr/doc_pub/Del3_1_V2.pdf).

This part of ISO 24615 proposes a metamodel for syntactic annotation together with a list of relevant data categories for syntactic annotation. The data categories are available on the ISOCat server (http://www.isocat.org/) in the syntax profile (as defined in ISO 12620:2009).

Language resource management — Syntactic annotation framework (SynAF) —

Part 1: Syntactic model

1 Scope

This part of ISO 24615 describes the syntactic annotation framework (SynAF), a high level model for representing the syntactic annotation of linguistic data, with the objective of supporting interoperability across language resources or language processing components. This part of ISO 24615 is complementary and closely related to ISO 24611 (MAF, morpho-syntactic annotation framework) and provides a metamodel for syntactic representations as well as reference data categories for representing both constituency and dependency information in sentences or other comparable utterances and segments.

2 Normative references

The following referenced documents are indispensable for the application 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 1087-1:2000, *Terminology work — Vocabulary — Part 1: Theory and application*

ISO 12620:2009, *Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources*

ISO 24611:2012, *Language resource management — Morpho-syntactic annotation framework*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 1087-1:2000, ISO 12620:2009, ISO 24611:2012 and the following apply.

3.1
adjunct
non-essential element associated with a verb as opposed to *syntactic arguments* (3.19)

Note 1 to entry: Adverbs are possible adjuncts for a sentence.

3.2
chunk
non-recursive *constituent* (3.4)

3.3
clause
group of *phrases* (3.14), usually containing a predicate

Note 1 to entry: A clause can be either a *main clause* (3.10) or a *subordinate clause* (3.17). In languages distinguishing finiteness, clauses whose predicate is a verb can be either finite or non-finite, depending on the form of the verb. A main clause alone can build a complete *sentence* (3.15). In the SynAF model, a clause is a special case of a *constituent* (3.4).

3.4
constituent
syntactic grouping of words [into *phrases* (3.14)], phrases [into *clauses* (3.3) or *other phrases*] or clauses [into a *sentence* (3.15)] on the base of structural (or hierarchical) properties

3.5
dependency
dependency relation
syntactic relation between *word forms* (3.24) or *constituents* (3.4) on the basis of the *grammatical functions* (3.7) that *constituents* play in relation to each other

3.6
syntactic edge
edge
triplet with a source *node* (3.12), a target *node*, and optional *annotations* (3.9)

Note 1 to entry: *Non-terminal nodes* (3.13) have an outgoing constituency syntactic edge.

3.7
grammatical function
grammatical role of a *wordform* (3.24) or *constituent* (3.4) within its embedding syntactic environment

Note 1 to entry: For example, a noun phrase (NP) can act as a subject within a *sentence* (3.15), or a noun may act as a subject dependent of a verb in a dependency graph. There is a grammatical relation between the subject – NP and the main verb in a sentence. All grammatical relations (subject – predicate, head – modifier, etc.) are subsumed under the concept of *dependency relations* (3.5), whether between terminal or non-terminal nodes.

3.8
syntactic head
head
part of a *constituent* (3.4) which determines its distribution (the syntactic environments in which the constituent may appear) and its grammatical properties (e.g. if the grammatical gender of the head is feminine, then the gender of the entire constituent will be feminine)

Note 1 to entry: The head of a constituent usually cannot be left out.

3.9
linguistic annotation
annotation
feature-value pair denoting a linguistic property of a linguistic segment

3.10
main clause
clause (3.3), which can act on its own as a complete *sentence* (3.15)

Note 1 to entry: In languages distinguishing finiteness, the main clause is usually finite. Example: *The train is late.*

3.11
modifier
part of a *constituent* (3.4) which ascribes a property to the *head* (3.8) of the *constituent*

Note 1 to entry: A modifier can be placed before or after the head of the *phrase* (3.14) (pre-modifier or post-modifier). Modifiers are optional in a constituent.

3.12
node
syntactic node
word form (3.24) or *constituent* (3.4) seen as an elementary syntactic component of a syntactic analysis

3.13
non-terminal node
syntactic node (3.12) which is not a *word form* (3.24)

Note 1 to entry: A non-terminal node has an outgoing constituency *edge* (3.6).

3.14
phrase
group of *word forms* (3.24) (usually containing one or more words) which can fulfill a *grammatical function* (3.7), e.g. in a *clause* (3.3)

Note 1 to entry: Empty phrases are permitted (being non-realised pronouns, sometimes marked as "pro", and having the role of subjects in clauses). A phrase is typically named after its *head* (3.8), for example noun phrases, verb phrases, adjective phrases, adverbial phrases and prepositional phrases. Phrases have been informally described as "bloated words", in that the parts of the phrase added to the head elaborate and specify the reference of the head. In our model, a phrase is a special case of a *constituent* (3.4).

3.15
sentence
related group of *word forms* (3.24) containing a predication, usually expressing a complete thought and forming the basic unit of discourse structure

Note 1 to entry: A sentence consists of one or more *clauses* (3.3). When describing speech, it is common to talk about "utterances" rather than sentences.

3.16
span

pair of points (p1, p2), where p1 ⩽ p2, identifying the segment of the document to which an *annotation* (3.9) is applied

Note 1 to entry: A multiple span is a sequence of spans where the ending point of each span is less than or equal to the starting point of the subsequent span.

3.17
subordinate clause
clause which fulfils a *grammatical function* (3.7) in a *phrase* (3.14) [for example a relative *clause* (3.3) modifying the *head* (3.8) noun of a nominal phrase] or in another clause

Note 1 to entry: A subordinate clause usually does not act on its own as a sentence, but is part of a larger sentence.

3.18
subcategorization frame
set of restrictions indicating the properties of the *syntactic arguments* (3.19) that can or must occur with a verb

EXAMPLE Alfred (/syntacticArgument/) reads a book (/syntacticArgument/) today (/adjunct/).

Note 1 to entry: The subject, indirect object and direct object are subcategorized *grammatical functions* (3.7) within a sentence; they are dependents of the verb (i.e. they can appear in subcategorization frames).

3.19
syntactic argument
functionally essential element that is required and given its interpretation by the head of its *phrase* (3.14) or the *node* (3.12) of which it is a dependent (e.g. the nominal argument of a prepositional phrase or verb)

Note 1 to entry: For verbs and verbal phrases, arguments identify the participants in the process referred to by the verb. In some frameworks, syntactic arguments are called complements.

3.20
syntactic graph
graph
connected set of *syntactic nodes* (3.12) and *edges* (3.6)

3.21
syntactic tree
syntactic graph (3.20) in which each node has a single parent

3.22
syntax
way in which *word forms* (3.24) are interrelated and/or grouped together into phrases, thus capturing the relations that exist between those units

3.23
terminal node
syntactic node (3.12) which is a single *word form* (3.24) or an empty element involved in a syntactic relation

3.24
word form
contiguous or non-contiguous entity from a speech or text sequence identified as an autonomous lexical item

4 SynAF metamodel

4.1 Introduction

Syntactic annotations have at least two functions in language processing:

a) to represent linguistic constituency, as in noun phrases (NP), describing a structured sequence of morpho-syntactically annotated items (including empty elements or traces generated by movements at the constituency level), as well as constituents built from non-contiguous elements, and

b) to represent dependency relations, such as head-modifier relations, and also including relations between categories of the same kind (such as the head-head relations between nouns in appositions, or nominal coordinations in some formalisms). The dependency information can exist between morpho-syntactically annotated items within a phrase (an adjective is the modifier of the head noun within an NP) or describe a specific relation between syntactic constituents at the clausal and sentential level (i.e. an NP being the "subject" of the main verb of a clause or sentence). The dependency relation can also be stated for empty elements (e.g. the pro element in romance languages, which serves a grammatical function).

As a consequence, syntactic annotations shall comply with a multi-layered annotation strategy interrelating syntactic annotation for both constituency and dependency as stated in the SynAF metamodel.

4.2 SynAF metamodel

4.2.1 Overview

The SynAF metamodel is represented as a set of UML classes complemented by UML attribute-value pairs, which represent the associated syntactic data categories. The SynAF textual descriptions specify more complete information about the SynAF classes, relations and extensions than can be included in the UML diagram. Developers shall define a data category selection (DCS) as specified for SynAF data category selection procedures (see Figure 1). The data categories given in Annex A shall be used for the representation of syntactic annotations.

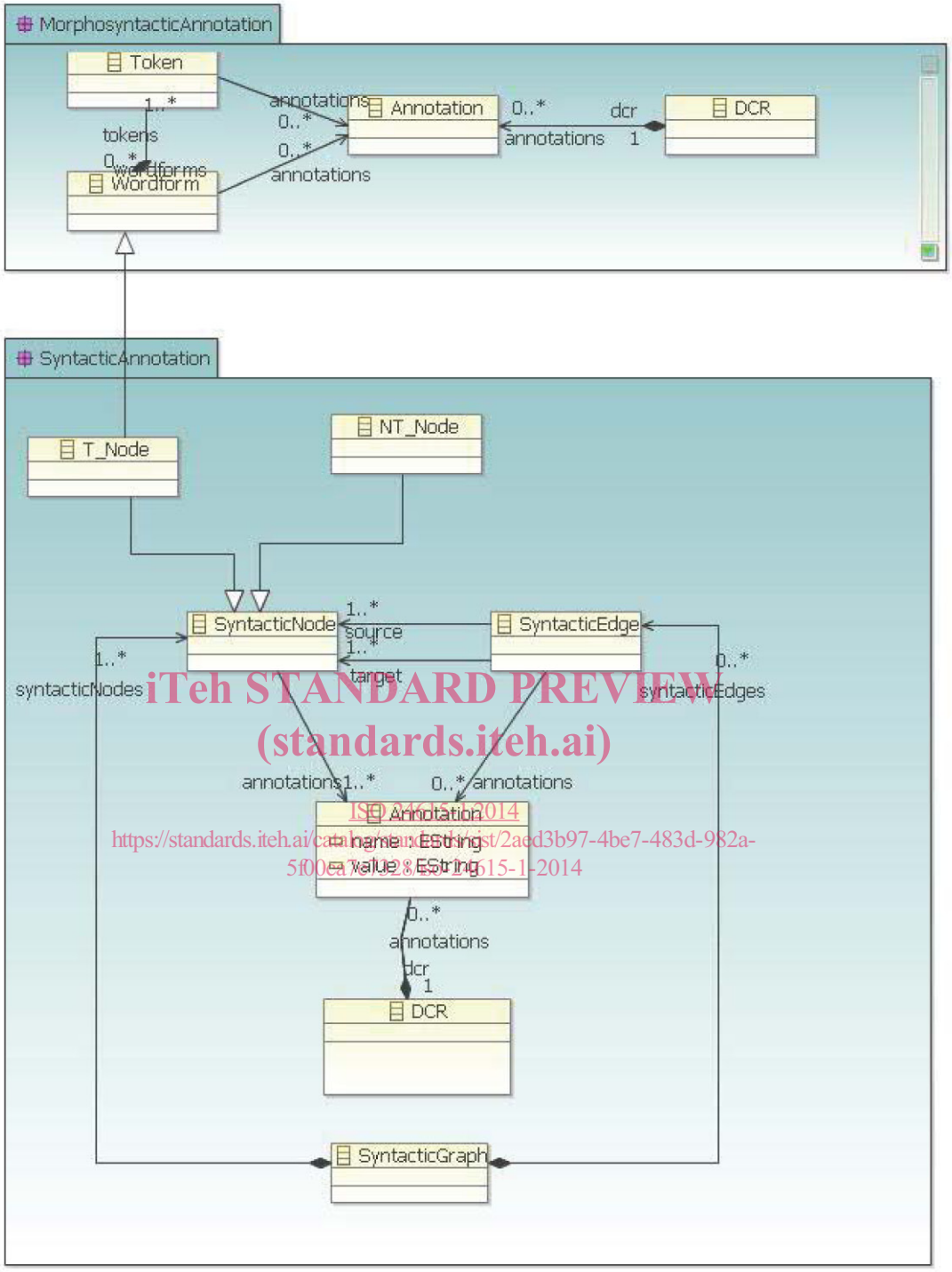

Figure 1 — SynAF metamodel (articulated with MAF)

4.2.2 SyntacticNode class

The *SyntacticNode* class is a generic class subsuming both the class of terminal nodes and the class of non-terminal nodes. Syntactic nodes can be involved in as many syntactic relations as necessary (see 3.6, **syntactic edges**).