International Standard

ISO 24611-1

First edition
2025-11

Language resource management — Morphosyntactic annotation framework (MAF) —

Part 1: Core model

Gestion des ressources linguistiques — Cadre d'annotation morphosyntaxique (MAF) —

Partie 1: Modèle de base

iTeh Standards
(https://standards.iteh.ai)
Document Preview

ISO 24611-1:2025
https://standards.iteh.ai/catalog/standards/iso/8650b33f-0e6e-4cc1-9a39-e585f15168c1/iso-24611-1-2025

Reference number
ISO 24611-1:2025(en)

© ISO 2025

iTeh Standards
(https://standards.iteh.ai)
Document Preview

ISO 24611-1:2025
https://standards.iteh.ai/catalog/standards/iso/8650b33f-0e6e-4cc1-9a39-e585f15168c1/iso-24611-1-2025

COPYRIGHT PROTECTED DOCUMENT

© ISO 2025

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
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

iTeh Standards
(https://standards.iteh.ai)
Document Preview
ISO 24611-1:2025
https://standards.iteh.ai/catalog/standards/sist/50b33f0e6e-4ee1-9a39-e585f15168e1/iso-24611-1-2025

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 document 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).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

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, *Language and terminology*, Subcommittee SC 4, *Language resource management*.

iTeh Standards
(https://standards.iteh.ai)
Document Preview

This first edition of ISO 24611-1 cancels and replaces ISO 24611:2012, which has been technically revised.

The main changes are as follows:

ISO 24611-1:2025
https://standards.iteh.ai/catalog/standards/iso/8650b33f-0e6e-4cc1-9a39-e585f15168c1/iso-24611-1-2025

— the data model is fully serialized in TEI XML;

— definitions and text have been revised;

— conformance conditions have been added;

— most of the former Clause 8, dealing with word lattices, has been removed and delegated to a planned ISO 24611-2;

— the annex of sample data categories has been removed in favour of an external repository of data categories.

A list of all parts in the ISO 24611 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.

© ISO 2025 – All rights reserved

Introduction

ISO/TC 37/SC 4 focuses on the definition of models and formats for the representation of annotated language resources. To this end, it has generalized the modelling strategy initiated by its sister committee, ISO/TC 37/SC 3, for the representation of terminological data (see Reference [21]), through which linguistic data models are seen as the combination of a generic data pattern (a metamodel), which is further refined through a selection of data categories that provide the descriptors for this specific annotation level. Such models are defined independently of any specific formats and ensure that an implementer has the necessary conceptual instrument with which to design and compare formats with regard to their degrees of interoperability.

One important aspect of representing any kind of annotation is the capacity to provide a clear and reliable semantics for the various descriptors used, either in the form of formal features and feature values, or directly as objects in a representation that is expressed, for instance, in XML. In order to be shared across various annotation schemas and encoding applications, such semantics should be implemented as a centralized repository of concepts: these concepts will henceforth be referred to as data categories. These data categories are envisioned as having the following two properties:

— From a technical point of view, they should provide unique, stable references (implemented as persistent identifiers, in the sense of ISO 24619) that specific encoding schemas can use to express their relatedness. By virtue of that, two annotations will be deemed equivalent if they are defined in relation to the same data categories (as feature and feature value).

— From a descriptive point of view, each unique semantic reference should be associated with precise documentation combining a full text elicitation of the meaning of the descriptor with the expression of specific constraints that bear upon the category.

iTeh Standards
(https://standards.iteh.ai)
Document Preview
ISO 24611-1:2025
https://standards.iteh.ai/catalog/standards/iso/8650b33f-0e6e-4cc1-9a39-e585f15168c1/iso-24611-1-2025

In the ISO 12620 series, a general framework for representing and maintaining such a repository of data categories has been developed, potentially encompassing all domains of language resources.

A possible instantiation of ISO 12620-1 is a “flat” marketplace of semantic objects, providing only a limited set of ontological constraints. The objective of such a setup would be to facilitate the maintenance of a comprehensive descriptive environment where new categories are easily inserted and re-used without the need for any strong consistency check with the repository at large. Indeed, the following kinds of constraints are part of the data category model, as defined in ISO 12620-1:

— Simple generic-specific relations, when these are useful for the proper identification of interoperability descriptors between data categories. For instance, the fact that /properNoun/ is a sub-category of /noun/ makes it possible to compare morphosyntactic annotations based on different descriptive levels of granularity.

— The description of conceptual domains that make it possible to identify, when known or applicable, the range of the possible values of so-called “complex data categories”. For instance, it can be used to record that possible values of /grammaticalGender/ (limited to a small group of languages, see Reference [21]), can be a subset of {/masculine/, /feminine/ and /neuter/}.

— Language-specific constraints, either in the form of specific application notes or as explicit restrictions bearing upon the conceptual domains of complex data categories. For instance, it is possible to express explicitly that /grammaticalGender/ in French can only take the two values: {/masculine/ and /feminine/}.

This document provides a comprehensive framework for the representation of morphosyntactic annotations (in their simplest form also referred to as “part of speech annotations” or “POS annotations”). This annotation level corresponds to the first lexical abstraction level over language data (textual or spoken) and, depending on the language to be annotated, as well as the characteristics of the annotation tool or annotation scheme that is being used, can vary enormously in structure and complexity.

In order to deal with such complex issues as ambiguity and determinism in morphosyntactic annotation, this document introduces a metamodel that draws a clear distinction between, on the one hand, the level of tokens (representing the surface segmentation of the source) and, on the other, the level of word-forms

© ISO 2025 – All rights reserved

(identifying lexical abstractions associated with groups of tokens). Both these levels can be represented as simple sequences and as local graphs in order to model constructions such as multiple segmentations and ambiguous compounds. Elements of these two levels can enter into any kind of *n*-to-*n* relationships.

As linguistic segments (sometimes called “markables” in the literature (see, for instance, Reference [18])), tokens can be delimited in the source document by means of inline mark-up, or they can be identified remotely (separately from the source document) by means of so-called “stand-off annotations”.

As linguistic abstractions, word-forms can be qualified by various linguistic features characterizing the morphosyntactic properties that are instantiated in the realization of the lexical entry within the annotated text. Such properties can range from the simple identification of a lemma up to an explicit reference to a lexical entry in a dictionary. In most existing applications of morphosyntactic annotation, linguistic properties are expressed by means of so-called “tags”. These codes refer to basic feature structures (see early examples in Reference [20]). Such codes can also provide morphological information, including its part of speech (e.g. noun, adjective, verb), and features such as number, gender, person, mood or tense.

In keeping with the general modelling strategy of ISO/TC 37, this document provides means of relating morphosyntactic tags expressed as feature structures (conforming to ISO 24610-1) to data categories (conforming to ISO 12620-1). Implementers are encouraged to use external reference taxonomies as described by ISO 12620-2 either directly, or by building on them in defining their own categories (appropriate in the coverage, scope or semantics to the requirements of the given encoding project), in conformity with ISO/TC 37 principles.

Associated to the metamodel, this document also provides a default XML syntax that can be used to serialize annotation models conforming to the morphosyntactic annotation framework (MAF). Since many existing projects are based on the Text Encoding Initiative (TEI) Guidelines, see Reference [31] (particularly in digital humanities, where a proper encoding of textual sources is essential), and since the TEI Guidelines already offer a variety of constructs and mechanisms to cope with many issues relevant to spoken corpora and their annotations (see Reference [22] and ISO 24624), the metamodel provided by this document is serialized as TEI XML. Many word-level annotation mechanisms used in this document elaborate on the proposal of Reference [23], implemented in the TEI Guidelines.

iTeh Standards
(https://standards.iteh.ai)
Document Preview

MAF consists of two parts, referred to as MAF Core (this document) and MAF Lattice (planned as ISO 24611-2).

Finally, this document forms the conceptual basis for the development of the ISO 24614 series on word segmentation, whereby all general principles and rules defined in ISO 24614-1, as well as the constraints expressed in additional parts for specific languages, can be understood according to the token versus word-form dichotomy.

ISO 24611-1:2025
https://standards.iteh.ai/catalog/standards/iso/24611-1-2025

© ISO 2025 – All rights reserved