SLOVENSKI STANDARD
oSIST ISO/DIS 24611-1:2024

01-oktober-2024

Upravljanje z jezikovnimi viri - Ogrodje za oblikoskladenjsko označevanje (MAF) - 1. del: Jedrni model

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

Ta slovenski standard je istoveten z: **ISO/DIS 24611-1**

ICS:

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

oSIST ISO/DIS 24611-1:2024 **en**

DRAFT International Standard

ISO/DIS 24611-1

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

ICS: 01.020

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

Secretariat: **KATS**

Voting begins on:
2024-07-25

Voting terminates on:
2024-10-17

This document is circulated as received from the committee secretariat.

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENTS AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

Reference number
ISO/DIS 24611-1:2024(en)

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

Foreword ..... **iv**

Introduction ..... **v**

1 Scope **1**

2 Normative references **1**

3 Terms and definitions **1**

4 The MAF metamodel **5**
4.1 Levels of description in the MAF metamodel 5
4.2 MAF in the standards landscape 6
4.3 Metadata 7
4.4 Structural ambiguities 7
4.5 MAF metamodel in detail 7

5 Token-level segmentation **9**
5.1 Introduction 9
5.2 Formal description: <seg> 9
5.3 Embedding notation 10
5.4 Stand-off notation 11
5.5 Normalization and script conversion 12
5.6 Inline token annotation strategies for token separation 13
5.6.1 Introduction 13
5.6.2 Adjacent tokens in embedded mode 13
5.6.3 Overlapping tokens 14

6 Word-forms as linguistic units **15**
6.1 Introduction 15
6.2 Formal description: <span> 16
6.3 Token attachment 16
6.3.1 One token : one word-form 16
6.3.2 Several contiguous tokens : one word-form 16
6.3.3 Several discontinuous tokens : one word-form 17
6.3.4 Zero token: one word-form 17
6.3.5 One token : several word-forms 18
6.4 Referencing lexical entries 18
6.5 Compound word-forms 19
6.6 Identification of word-forms 20

7 Morphosyntactic content **20**
7.1 Introduction 20
7.2 Using feature structures 20
7.3 Compact morphosyntactic tags 21
7.4 FSR libraries 21
7.5 Designing morphosyntactic tagsets 23

8 Handling ambiguities **24**
8.1 Introduction 24
8.2 Word-form content ambiguities 24
8.3 Lexical and structural ambiguities 25

9 Conformance **25**

Annex A (informative) **Examples** **26**

Annex B (informative) **Referencing externally defined data categories** **31**

Bibliography **34**

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).

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*.

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

The main changes are as follows:

- the data model is fully serialised in TEI XML;
- definitions and text have been revised;
- conformance conditions have been added;
- most of Clause 8, dealing with word lattices, has been removed and delegated to a planned part 2 of the ISO 24611 series;
- informative 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.

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 [22]), 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: we will henceforth refer to these concepts 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.

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. That initiative makes it possible to implement an online environment providing access to data categories that various language resource-related activities within ISO should align against.

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, in the sense of the ISO/IEC 11179 series, to identify, when known or applicable, the possible value 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 [22]), could 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’ or ‘POS’). 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 (identifying lexical abstractions associated with groups of tokens). Both these levels can be represented as simple sequences and as local graphs 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[19],), *tokens* may be delimited in the source document by means of inline mark-up, or they may 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 characterising the morphosyntactic properties that are instantiated in the realization of the lexical entry within the annotated text. Such properties may 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[21]). Such codes may also provide morphological information, including its part of speech (e.g. noun, adjective or verb), and features such as number, gender, person mood and verbal tense.

In keeping with the general modelling strategy of ISO/TC 37, this document provides means of relating morphosyntactic tags expressed as feature structures (compliant with ISO 24610-1) to data categories (compliant with ISO 12620-1). Implementers are encouraged to use external reference taxonomies as described by ISO 12620-1 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 compliance 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 compliant with the Morphosyntactic Annotation Framework (MAF). Since many existing projects are based on the Text Encoding Initiative (TEI) guidelines (see [32]) — 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 [23] and ISO 24624), the metamodel provided by this document is serialized as TEI XML. Many word-level annotation mechanisms used here elaborate on the proposal of Reference [24], implemented in the TEI Guidelines.

Finally, it should be noted here that 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, are to be understood according to the token vs. word-form dichotomy.

Language resource management — Morphosyntactic annotation framework (MAF) —

Part 1:
Core model

1 Scope

This document provides a framework for the representation of annotations of word-sized units in texts. Such annotations describe tokens, their relationship with lexical units (word-forms), and the relevant morphosyntactic properties. This document proposes a metamodel for morphosyntactic annotation that can be augmented with references to data categories contained in an ISO-12620-1-compliant data category repository. It also defines an XML serialization for morphosyntactic annotations, according to the principles laid out in the TEI Guidelines.

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

Structural ambiguities are not in the scope of this document, and neither is the structure and composition of morphosyntactic tagsets.

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 24610-1:2006, *Language resource management — Feature structures — Part 1: Feature structure representation*

TEI P5, Guidelines for Electronic Text Encoding and Interchange. Version 4.7.0. Last updated on 16 November 2023. TEI Consortium. https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html

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

3 Terms and definitions

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

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

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

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

3.1
morphology
description of the structure and formation of *words* (3.7)

Note 1 to entry: Morphology is traditionally divided into (a) *word-formation* (3.2) – dealing with the formation of complex *lexemes* (3.6) out of simpler lexemes: by means of derivation (often signalled by affixation, i.e., addition of a *morpheme* (3.11)) or by means of compounding (combining two or more lexemes), and into (b) *inflection* (3.3) that creates *inflected forms* (3.4).

3.2
word-formation
branch of *morphology* (3.1), dealing with the creation of new *lexemes* (3.6) by the processes of derivation and compounding

3.3
inflection
branch of *morphology* (3.1), dealing with contextual realizations of *lexemes* (3.6) as *inflected forms* (3.4)

3.4
inflected form
concrete form that a *lexeme* (3.6) can take when used in a sentence or a phrase

3.5
word-form
abstract instantiation of a *lexeme* (3.6) with the values of *morphosyntactic features* (3.12) fixed in a syntactic context

Note 1 to entry: Word-forms may have no acoustic or graphic realization, or may correspond to one or more *tokens* (3.21), not necessarily forming a contiguous sequence.

3.6
lexeme
abstract, fundamental unit in the lexicon of a language, comprising semantic, formal (phonetic and/or graphemic) and grammatical information

Note 1 to entry: A complex lexeme is the result of *word-formation* (3.2) (derivation or compounding) processes; a simple lexeme can be thought of as the base for such processes. In a *lexical entry* (3.9), a lexeme is identified by a *lemma* (3.8). *Word-forms* (3.5) are results of the interaction of lexemes with the grammatical system of the given language.

3.7
word
lexeme (3.6), *word-form* (3.5) or *token* (3.21)

Note 1 to entry: The term *word* is notoriously ambiguous, standing (at least) for lexeme, word-form or token, depending on the context of its use. This document attempts to disambiguate this term where relevant.

3.8
lemma
conventional form chosen to represent a *lexeme* (3.6)

Note 1 to entry: In European languages, the lemma is usually the *singular* if there is a variation in number, the *masculine* form if there is a variation in gender, and the *infinitive* for all verbs. In some languages, certain nouns are defective in the singular form; in these cases, the plural is chosen. For verbs in Arabic, the lemma is usually deemed to be the third person singular with the accomplished aspect.

Note 2 to entry: The term *lemma* is most often used in the context of corpora, as a device to capture the identity of *tokens* (3.21) and establish basic correspondence between a token and a *lexical entry* (3.9). The term that corresponds to *lemma* in the context of lexicons is *headword*. Mismatches between the two are possible due to the varying macro- and microstructure of lexical entries. In order to handle such mismatches, apart from lemmas, direct references to dictionary entries are sometimes added to tokens or *word-forms* (3.5) in corpora.

3.9
lexical entry
container for managing a set of *word-forms* (3.5) and possibly one or more meanings that describe a *lexeme* (3.6)

3.10
lexicon
resource comprising a collection of *lexical entries* (3.9) for a language

3.11
morpheme
exponent that signals a modification of a *lexeme* (3.6)

Note 1 to entry: This definition adheres to a lexeme-based approach to morphology where it is the lexeme, not the morpheme, that encodes the linguistic sign. On this approach, the morpheme is a unit of form (an exponent) that marks various kinds of modifications (e.g. derivation or inflection) of a lexeme.

Note 2 to entry: Morphemes can usually be divided into derivational and inflectional (signalling a morphosyntactic category); sometimes a modification of a lexeme is not overtly marked, and sometimes the morpheme is a combined (fused) exponent of various kinds of morphosyntactic information.

Note 3 to entry: On morpheme-based (as opposed to lexeme-based) approaches, the morpheme is defined as the minimal linguistic sign (a combination of the meaning and the form). On these approaches, the term *morph* is used roughly in the meaning that is used for the term *morpheme* in this standard.

3.12
morphosyntactic feature
feature induced from either the *inflected form* (3.4) of a *lexeme* (3.6) or from its syntactic context, or both

EXAMPLE "grammaticalGender"

Note 1 to entry: Universal Dependencies (see [27]) offer a set of general and language-specific features and values, designed for pragmatically uniform cross-linguistic grammatical description.

3.13
part of speech
POS
grammatical category
category assigned to a *word* (3.7) based on its grammatical and semantic properties

EXAMPLE Noun, verb.

3.14
morphosyntactic tag
tag
label identifying a *feature structure* (3.16) used to qualify a *word-form* (3.5) within an established taxonomy

Note 1 to entry: Morphosyntactic tags can be atomic labels ("N" for 'noun'), but very often they are mnemonic representations for the feature structures that they identify ("NNL2" for 'plural locative noun' in the CLAWS-7 *tagset* (3.15), see [29]). The relevant feature structures can also be encoded by character vectors, as in "N12201" for 'common noun, feminine, plural, countable' in the EAGLES intermediate tagset (see [30]) or by agglutinated shorthand feature identifiers, as in "subst:pl:gen:m3" for 'noun, plural, genitive, masculine, inanimate' in the NKJP tagset (see [31]).

3.15
morphosyntactic tagset
tagset
comprehensive set of *morphosyntactic tags* (3.14) used for the morphosyntactic description of a language

3.16
feature structure
set of *feature specifications* (3.17)

[SOURCE: ISO 24610-1:2006, 3.10, modified: Note removed]

3.17
feature specification
assignment of a value to a feature

Note 1 to entry: Formally, it is treated as a pair of a feature and its value.

[SOURCE: ISO 24610-1:2006, 3.9]

3.18
phoneme
minimal unit in the sound system of a language

3.19
phonetic transcription
representation or modelling of spoken language based on the sound system of the respective language

[SOURCE: ISO 24624:2016, 3.5]

3.20
character
element of a writing system, whether or not alphabetical, that represents a *phoneme* (3.18), a syllable, a *word* (3.7) or even prosodic characteristics of the language, by using graphical symbols (letters, diacritical marks, syllabic signs, punctuation marks, prosodic accents, etc.) or a combination of these signs (a letter having an accent or a diacritical mark)

EXAMPLE a, B, ω or Γ are, therefore, characters as well as basic letters.

Note 1 to entry: See also ISO/IEC 2382:2015, 2121335.

[SOURCE: ISO 7098:2015, 2.1]

3.21
token
non-empty contiguous sequence of *characters* (3.20) in a document

Note 1 to entry: For editorial reasons, some annotation schemes may extend the notion of token to an empty sequence.

3.22
tokenization
process that segments a language data stream into individual tokens (3.21)

3.23
transcription
general transcription
form resulting from a type of *script conversion* (3.25) whereby *characters* (3.20) of one *script* (3.26) are mapped onto characters of another script)

3.24
transliteration
form resulting from the *conversion* (3.25) of one *script* (3.26) into another, usually through a one-to-one correspondence between *characters* (3.20)

3.25
script conversion
transcription (3.23) and *transliteration* (3.24)

[SOURCE: ISO 5127:2017, 3.1.6.13]