INTERNATIONAL STANDARD

ISO 24611

First edition
2012-11-01

Language resource management — Morpho-syntactic annotation framework (MAF)

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

Reference number
ISO 24611:2012(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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

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.

ISO 24611 was prepared by Technical Committee ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management*.

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 generalised the modelling strategy initiated by its sister committee, SC 3, for the representation of terminological data [Romary, 2001], through which linguistic data models are seen as the combination of a generic data pattern (a meta-model), 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 a semantics should be implemented as a centralised registry of concepts: we will henceforth refer to these as data categories. As such, data categories should bear the following constraints.

- From a technical point of view, they must provide unique, stable references (implemented as persistent identifiers, in the sense of ISO 24619) such that the designer of a specific encoding schema can refer to them in his or her specification. By doing so, two annotations will be deemed to be equivalent when they are in fact 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 recent years, ISO has developed a general framework for representing and maintaining such a registry of data categories, encompassing all domains of language resources. This initiative, described in ISO 12620, has led to the implementation of an online environment providing access to all data categories that have been standardized in the context of the various language resource-related activities within ISO, or specifically as part of the maintenance of the data category registry. It also provides access to the various data categories that individual language technology practitioners have defined in the course of their own work and decided to share with the community.

The ISO data category registry, as available through the ISOCat (www.isocat.org) implementation, is intended as a ‘flat’ marketplace of semantic objects, providing only a limited set of ontological constraints. The objective there is to facilitate the maintenance of a comprehensive descriptive environment where new categories are easily inserted and reused without the need for any strong consistency check with the registry at large. Indeed, the following basic constraints are part of the data category model, as defined in ISO 12620:

- 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 morpho-syntactic annotations based on different descriptive levels of granularity;

- the description of conceptual domains, in the sense of ISO 11179, 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 [Romary 2011]), could be a subset of {/masculine/, /feminine/ and /neutral/};

- 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 International Standard provides a comprehensive framework for the representation of morpho-syntactic (also referred to as part-of-speech) annotations. Such an annotation level corresponds to a first lexical abstraction level over language data (textual or spoken) and, depending on the language to be annotated, together with 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 morpho-syntactic annotation, this International Standard introduces a meta-model that draws a clear distinction between the two levels of tokens (representing the surface segmentation of the source) and word-forms (identifying lexical abstractions associated with groups of tokens). These two levels share the following specificities: on the one hand, they can be represented as simple sequences and as local graphs such as multiple segmentations and ambiguous compounds; on the other hand, any n-to-n combination can stand between word forms and tokens.

As linguistic segments (sometimes called ‘markables’ in the literature [see, for instance, Carletta et al. 1997]), *tokens* may be embedded in the source document as inline mark-up, or they may point remotely to it by means of so-called stand-off annotations.

As linguistic abstractions, *word-forms* can be qualified by various linguistic features characterising the morpho-syntactic properties that are instantiated in the realisation of the lexical entry within the annotated text. Such properties may range from the simple indication of a lemma up to an explicit reference to a lexical entry in a dictionary. In most existing applications of morpho-syntactic annotation, linguistic properties are expressed by means of so-called tags; these codes refer to basic feature structures (see early examples in Monachini and Calzolari, 1994). 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 International Standard/MAF provides means of relating morpho-syntactic tags expressed as feature structures (compliant with ISO 24610) to the data categories available in ISOCat. A normative annex of this International Standard elicits a core set of data categories that can be used as reference for most current morpho-syntactic annotation tasks in a multilingual context. However, when implementers of this International Standard find these categories inappropriate in either coverage, scope or semantics, they are encouraged to use ISOCat to define their own categories in compliance with ISO/TC 37 principles.

Associated to the meta-model, MAF also provides a default XML syntax that may be used to serialise MAF-compliant annotation models. Since many existing projects are based on the text encoding initiative (TEI) guidelines (www.tei-c.org) — particularly in digital humanities, where a proper encoding of textual sources is essential — this International Standard will also provide clues about how to articulate the MAF model with TEI-compliant encodings. Indeed, the TEI guidelines already offer a variety of constructs and mechanisms to cope with many issues relevant to spoken corpora and their annotations (Romary and Witt, 2012).

Finally, it should be noted here that this International Standard 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–word-form dichotomy.

Language resource management — Morpho-syntactic annotation framework (MAF)

1 Scope

This International Standard provides a framework for the representation of annotations of word-forms in texts; such annotations concern tokens, their relationship with lexical units, and their morpho-syntactic properties.

It describes a metamodel for morpho-syntactic annotation that relates to a reference to the data categories contained in the ISOCat data category registry (DCR, as defined in ISO 12620). It also describes an XML serialization for morpho-syntactic annotations, with equivalences to the guidelines of the TEI (text encoding initiative).

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

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 24610-1 and the following apply.

3.1
DAG
directed acyclic graph
graph with directed edges and no cycles

Note 1 to entry: DAGs are a subset of *finite state automata* (3.4).

3.3
feature structure
set of feature specifications, used in the morpho-syntactic annotation framework (MAF) to express morpho-syntactic content

Note 1 to entry: Feature structures are described in ISO 24610-1.

3.4
FSA
finite state automata
graphs made up of states with an initial state and a final state, and a finite set of transitions from state to state

Note 1 to entry: See also *DAG* (3.1).