SLOVENSKI STANDARD
SIST-TP ISO/TR 21636-2:2024

01-november-2024

Jezikovno kodiranje - Ogrodje za jezikovne različice - 2. del: Opis ogrodja

Language coding — A framework for language varieties — Part 2: Description of the framework

Identification et description des variétés de langues — Partie 2: Description

Ta slovenski standard je istoveten z: **ISO/TR 21636-2:2023**

ICS:

01.140.20	Informacijske vede	Information sciences

SIST-TP ISO/TR 21636-2:2024   **en**

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

TECHNICAL
REPORT

ISO/TR 21636-2

First edition
2023-03

Language coding — A framework for language varieties —

Part 2:
Description of the framework

Identification et description des variétés de langues —

Partie 2: Description

ISO/TR 21636-2:2023(E)

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

COPYRIGHT PROTECTED DOCUMENT

ISO/TR 21636-2:2023(E)

Contents

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

ISO/TR 21636-2:2023(E)

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 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 2, *Terminology workflow and language coding*.

A list of all parts in the ISO 21636 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

More and more digital language resources (LRs) are being created (also by retro-digitization), archived, processed and analysed. In this context, detailed and exact characterization of language varieties present in a given language use event is quickly gaining importance. Here, language use includes all modalities such as written, spoken or signed, and also new forms of language use supported by digital technology (in social media and similar forms of digital communication). But this is just one way in which languages vary internally. Others include, for instance, the well-known regional (dialectal) and social variation.

While in the past a primary goal of working with LRs was the archiving and preservation of LRs, new goals have emerged and are still emerging:

— institutions and individuals need to exchange metadata (that is, bibliographic description data and other secondary information) for making the information on existing LRs widely available in a harmonized form;

— researchers are looking for the primary data (that is, the LRs themselves) for many different research purposes, including research on linguistic variation;

— researchers and developers need LRs for the development of more advanced language technologies (LTs) and for testing purposes, as LTs, in particular speech recognition and language analysis, are entering more and more dimensions of human communication.

In order to achieve the above-mentioned goals and purposes, along with others not outlined in this document, a standardized set of metadata for the identification of language varieties is important to guarantee frictionless exchange of secondary information. Well-organized metadata also help to indicate the degree of interoperability (equalling re-usability and re-purposability of LRs), and the applicability of LTs to different situations or LRs over time. These metadata are applicable in eBusiness, eHealth, eGovernment, eInclusion, eLearning, smart environments, ambient assisted living (AAL) and virtually all other applications which depend on information about LRs. A clear metadata approach is also a prerequisite for the durability of language resource archiving (in particular in the case of cultural heritage and scientific research data).

The identification of different individual languages is the subject of ISO 639[1]), which identifies existing (living, extinct and historical) individual languages, as well as language groups. This document, and the ISO 21636 series in general, presupposes and complements ISO 639 by extending the language code framework in order to allow for the identification of language varieties of different types (such as geographical, social and modal varieties, among others). The identification of language varieties can then be included in general, library and archival metadata for describing LRs (which can also include technical information, time and location of recording, and similar general information, which are not part of the ISO 21636 series).

The provisions of the ISO 21636 series cover:

— a general conceptual framework to deal coherently with language-internal linguistic variation;

— general rules for the identification and description of language varieties;

— a set of dimensions and open-ended or closed lists of values that can be assigned to each respective dimension;

— a set of metadata categories and examples for the respective possible values, grouped according to the most important aspects of the description of events of language use and resulting LRs, related to linguistic variation.

The metadata categories and values addressed in this document can be candidates for a future highly granular coding of language varieties based on these comprehensive principles. Thus, this document

---

1)   Under preparation. Status at the time of publication: ISO/FDIS 639:2023.

ISO/TR 21636-2:2023(E)

(and the ISO 21636 series in general) conforms to the "recommendations on software and content development principles 2010", and fits within the general framework of the ISO/IEC 11179 series for metadata.

Stakeholders include, but are not limited to:

— information and communication technologies (ICTs) industry (including LTs);

— libraries;

— the media industry (including entertainment);

— internet communities;

— people engaging in language documentation and preservation;

— language archivists;

— translators and interpreters;

— researchers (linguists, in particular sociolinguists, ethnologists, sociologists, etc.);

— people and institutions providing language training;

— emerging new user communities.

It is anticipated that these stakeholders need to refer not only to a certain individual language, but also to a certain language variety, for instance for oral human-computer interaction, or for tailoring a certain LR or tool to the needs and specific environment of a target user group. In order to identify the dimension(s) of linguistic variation internal to individual languages involved, and the respective relevant language varieties, a first step is to achieve the needed specificity. Adapting a conceptually sound, uniform framework of reference as developed in this document is superior to the proliferation of different individual ad hoc solutions.

Language coding — A framework for language varieties —

Part 2:
Description of the framework

1 Scope

This document, and the ISO 21636 series in general, provides the general principles for the identification and description of varieties of individual human languages. It, therefore, does not apply to:

— artificial means of communication with or between machines such as programming languages;

— those means of human communication which are not fully or largely equivalent to human language such as individual symbols or gestures that carry isolated meanings but cannot be freely combined into complex expressions.

This document together with the other parts of the ISO 21636 series establishes the dimensions of linguistic variation as well as core values necessary to identify individual varieties in these dimensions or sub-dimensions.

This document forms the basis for the other parts by outlining the general framework for language varieties.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

No terms and definitions are listed in this document.

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/

4 Linguistic variation and language varieties

4.1 Linguistic variation

Individual human languages differ from one another, and variation exists within each one. Since language variation is inherent, this variation is also present in LRs. This document covers the description of LRs that represent instances of use of individual languages regarding their status with respect to linguistic variation.

While individual languages originally emerged and were mainly used for communication between humans, their use is increasingly supported by ICTs. Events of language use involving machines are also covered by this document.

Within individual human languages, linguistic variation occurs in distinct dimensions (listed in 4.2 and described in detail in the remaining text), resulting in different kinds of language varieties. Each of the dimensions is independent from the others, although mutual influences exist. Each linguistic

manifestation in a given individual language, such as a written text, an utterance, an entry in a lexical database, etc., can, therefore, be characterized by its location in each of these dimensions of linguistic variation.

Some language varieties can be further differentiated into sub-varieties, but this level of detail is not the focus of this document.

Furthermore, the borders between individual languages can be hard to establish, especially when different groups of people disagree in their assessment of the mutual intelligibility or the socio-political situation so that the status of a given set of idiolects as an individual language, language variety or group of related languages is disputed. This difficulty is not addressed in this document (see ISO 639).

Similarly, the borders of language varieties are sometimes hard to establish. In many cases, different language varieties being distinguished on the same dimension of linguistic variation (for instance different dialects) do overlap — in other words, there can be idiolects that fulfil all respective criteria to belong to both varieties. This is particularly true for the time dimension (see 4.4). However, it does not invalidate the establishment of varieties.

In addition, individual languages and their language varieties are under constant gradual, sometimes rapid, change so that the following applies:

— A given language variety can over time gain the status of an individual human language (to receive its own language identifier in ISO 639). For instance, different dialects of Vulgar Latin developed into distinct Romance languages. Conversely, over time an individual language can develop into a language variety of another individual language, as appears to have been the case with a postulated Spanish-based creole language spoken in Cuba that would have developed into a variety of Cuban Spanish. In this kind of case, the resulting variety would be subject to the framework described in this document, in which case an existing ISO 639 language identifier assigned to the earlier form of this language would come to designate an extinct individual language.

— Linguistic expressions and features belonging to a given language variety at one point can shift along one or more dimensions of linguistic variation (for instance markedly informal expressions can become acceptable even in formal contexts, or regional expressions can spread over large parts of the geographical territory where the individual language is spoken).

This document focuses on a framework for the identification and description of linguistic variation within a language, that is, of language varieties and sub-varieties. It does not attempt to document all varieties of all individual languages, rather to list the types of descriptors needed to precisely and fully characterize the linguistic variety represented by an utterance or LR. It does so by describing the dimensions of variation within languages, and by indicating the major language variations that typically occur within individual languages.

This document considers individual languages to be sets of idiolects, which in turn are sets of expressions of human language with individual meanings. Each idiolect is characterized by a coherent system of structural features, and is homogenous by definition. This understanding of "idiolect" should not be confused with its use in the sense of "personal variety" — the total repertoire a speaker uses in a language, which is heterogenous, varying according to different situations and/or media, and thus unsuited to defining the smallest unit of linguistic variation.

In this framework, all language varieties are subsets of individual languages. Each individual language is classified into different varieties according to different kinds of external criteria and, at the same time, of structural criteria (see below). The resulting language varieties in each classification can overlap, and they can be sub-classified into smaller language varieties, again according to different criteria.

Each language variety is characterized by structural criteria, in particular, of the sound system (phonetics and phonology) or its counterpart in other language modalities (the visual-spatial system, or system of graphemes, etc.), the morphology and syntax, the lexicon and the semantic system. At the same time, each language variety is also characterized by certain external properties; for example, to be used in a certain region, at a certain time, in a formal situation. These criteria are organized into a