PD ISO/TR 21636-2:2023



BSI Standards Publication

Language coding — A framework for language varieties

Part 2: Description of the framework

bsi.

National foreword

This Published Document is the UK implementation of ISO/TR 21636-2:2023.

The UK participation in its preparation was entrusted to Technical Committee TS/1, Terminology.

A list of organizations represented on this committee can be obtained on request to its committee manager.

Contractual and legal considerations

This publication has been prepared in good faith, however no representation, warranty, assurance or undertaking (express or implied) is or will be made, and no responsibility or liability is or will be accepted by BSI in relation to the adequacy, accuracy, completeness or reasonableness of this publication. All and any such responsibility and liability is expressly disclaimed to the full extent permitted by the law.

This publication is provided as is, and is to be used at the recipient's own risk.

The recipient is advised to consider seeking professional guidance with respect to its use of this publication.

This publication is not intended to constitute a contract. Users are responsible for its correct application.

This publication is not to be regarded as a British Standard.

© The British Standards Institution 2023
Published by BSI Standards Limited 2023

ISBN 978 0 539 04971 8

ICS 01.140.20

Compliance with a Published Document cannot confer immunity from legal obligations.

This Published Document was published under the authority of the Standards Policy and Strategy Committee on 30 April 2023.

Amendments/corrigenda issued since publication

Date	Text affected

PD ISO/TR 21636-2:2023

TECHNICAL REPORT

ISO/TR 21636-2

First edition
2023-03-23

Language coding — A framework for language varieties —

Part 2: Description of the framework

Identification et description des variétés de langues —

Partie 2: Description

Reference number
ISO/TR 21636-2:2023(E)

© ISO 2023



COPYRIGHT PROTECTED DOCUMENT

© ISO 2023, Published in Switzerland

All rights reserved. Unless otherwise specified, 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
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

Contents

Page

Foreword ..... **iv**
Introduction ..... **v**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Linguistic variation and language varieties **1**
4.1 Linguistic variation 1
4.2 Dimensions of linguistic variation 3
4.3 The space dimension and its varieties 4
4.4 The time dimension and its varieties 5
4.5 The social group dimension and its varieties 5
4.6 The medium dimension and its varieties 6
4.7 The situation dimension and its varieties 8
4.8 The individual speaker dimension and its varieties 10
4.9 The proficiency dimension and its varieties 10
4.10 The communicative functioning dimension and its varieties 12
Bibliography **14**

© ISO 2023 – All rights reserved

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.

© ISO 2023 – All rights reserved

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 2023 – All rights reserved

(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.

© ISO 2023 – All rights reserved

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

© ISO 2023 – All rights reserved