



FINAL DRAFT International Standard

ISO/FDIS 21636-1

Language coding — A framework for language varieties —

Part 1: Vocabulary

ISO/TC 37/SC 2

Secretariat: **SCC**

Voting begins on:
2024-03-11

Voting terminates on:
2024-05-06

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

[ISO/FDIS 21636-1](#)

<https://standards.iteh.ai/catalog/standards/iso/782a738d-7020-4e46-8264-e08f35413f23/iso-fdis-21636-1>

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.

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.

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

[ISO/FDIS 21636-1](https://standards.iteh.ai/catalog/standards/iso/782a738d-7020-4e46-8264-e08f35413f23/iso-fdis-21636-1)

<https://standards.iteh.ai/catalog/standards/iso/782a738d-7020-4e46-8264-e08f35413f23/iso-fdis-21636-1>



COPYRIGHT PROTECTED DOCUMENT

© ISO 2024

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

Foreword	iv
Introduction	v
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
3.1 Terms related to language and languages.....	1
3.2 Terms related to linguistic variation and language varieties.....	3
3.3 Terms related to dimensions of linguistic variation.....	4
3.4 Terms related to types of language varieties.....	5
3.5 Terms related to specific language modalities.....	8
3.6 Terms related to major language registers.....	11
3.7 Terms related to the documentation of language resources.....	12
3.8 Terms related to certainty.....	13
Bibliography	14
Index	15

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

[ISO/FDIS 21636-1](https://standards.iteh.ai/catalog/standards/iso/782a738d-7020-4e46-8264-e08f35413f23/iso-fdis-21636-1)

<https://standards.iteh.ai/catalog/standards/iso/782a738d-7020-4e46-8264-e08f35413f23/iso-fdis-21636-1>

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 2, *Terminology workflow and language coding*.

A list of all parts of 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.

<https://standards.iteh.ai/catalog/standards/iso/782a738d-7020-4e46-8264-e08f35413f23/iso-fdis-21636-1>

Introduction

An increasing amount of digital language resources (LRs) are being created (including via retro-digitization), archived, processed and analysed. Within this context, the 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). Such modalities demonstrate one way in which languages vary internally. Others include, for instance, familiar regional (dialectal) and social variation.

In the past, a primary goal of working with LRs was the archiving and preservation of LRs. However, new goals have now emerged and are still emerging:

- Institutions and individuals need to exchange metadata (i.e. bibliographic description data and other secondary information) for making the information on existing LRs widely available in a harmonized form.
- Researchers are identifying primary data (i.e. the LRs themselves) for various 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, because LTs, in particular those concerning speech recognition and language analysis, are entering more dimensions of human communication.

In order to achieve the above-mentioned goals and purposes, along with others not outlined in the ISO 21636 series, a standardized set of metadata for the identification of language varieties is important for guaranteeing the 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 information-rich applications which depend on information about LRs. A clear metadata approach is also a prerequisite for the durability of LR archiving (in particular in the case of cultural heritage and scientific research data).

ISO 639 provides a framework for identifying the individual languages used in an LR. The ISO 21636 series presupposes and complements ISO 639 in that it extends the language coding framework in order to allow for the identification of different types of language varieties (e.g. geographical, social, modal). The identification of language varieties can then be included in general metadata, library metadata and archival metadata for describing LRs (which may also include technical information, time and location of recording, and similar general information, which are not included in the ISO 21636 series).

The conceptual framework developed in this document for dealing with linguistic variation respects the major approaches represented in the linguistic literature without simply reproducing them. The framework is closest though in general orientation and in a number of details, such as the role assigned to idiolects, to work of a type represented by Lieb^[5].

This document comprises:

- terms and definitions underlying a general conceptual framework to coherently deal with language-internal linguistic variation;
- terms and definitions for a set of dimensions for identifying and describing language varieties.

Stakeholders include, but are not limited to:

- information and communication technologies (ICTs) industry (including LTs);
- libraries;
- the media industry (including entertainment);
- internet communities;

ISO/FDIS 21636-1:2024(en)

- people engaging in language documentation and preservation;
- language archivists;
- researchers (linguists, in particular sociolinguists, ethnologists, sociologists, etc.);
- people and institutions providing language training;
- emerging new user communities.

It is anticipated that these stakeholders will 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 LT to the needs and specific environment of a target user group. An initial step towards achieving the needed specificity involves the ability to identify the dimension(s) of linguistic variation internal to individual languages involved, and the respective relevant language varieties. A conceptually sound uniform framework of reference as developed in the ISO 21636 series is superior to the proliferation of different individual ad-hoc solutions.

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

[ISO/FDIS 21636-1](https://standards.iteh.ai/catalog/standards/iso/782a738d-7020-4e46-8264-e08f35413f23/iso-fdis-21636-1)

<https://standards.iteh.ai/catalog/standards/iso/782a738d-7020-4e46-8264-e08f35413f23/iso-fdis-21636-1>

Language coding — A framework for language varieties —

Part 1: Vocabulary

1 Scope

The ISO 21636 series provides a framework for the identification and description of varieties of all individual human languages (see ISO 639).

It is applicable to sign languages.

It does not apply to:

- artificial means of communication with or between machines (such as programming languages);
- those means of human communication which are neither fully nor largely equivalent to human language (such as sets of individual symbols or gestures that each carry isolated meanings but cannot be freely combined into complex expressions).

This document defines the terms necessary to identify basic dimensions and sub-dimensions of linguistic variation and the resulting varieties, including major modalities of human communication.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

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 Terms related to language and languages

3.1.1

human language

means of communication characterized by a systematic use of sounds, visual-spatial signs, characters or other written symbols or signs that can be combined to express or communicate meaning or a message between humans

Note 1 to entry: Human language was originally developed for, and mainly used in, direct communication between humans. Today its use is increasingly supported by information and communication technologies (ICTs).

Note 2 to entry: As the term “language” can represent different concepts, it is not listed as a synonym to the term “human language”.

Note 3 to entry: Visual-spatial signs are indicated under *signed modality* (3.5.4).

3.1.2**idiolect**

comprehensive set of all expressions of *human language* (3.1.1) with their meaning, characterized by a coherent system of structural features, which is capable of coding complex facts and thoughts, potentially used by a given individual person, in a given type of situation, at a given time, and in a given medium

Note 1 to entry: Typically, a person has command of several idiolects of an *individual language* (3.1.3), for instance written and spoken idiolects (belonging to different language modalities; see 3.4.7 and 3.5), and idiolects for situations with different degrees of formality (belonging to different language registers; see 3.4.8 and 3.6).

3.1.3**individual language**

individual human language

largest set of *idiolects* (3.1.2), used by different *speakers* (3.1.5), which are all interconnected through high mutual intelligibility, or through a chain of high mutual intelligibility, or which are sociopolitically considered as a unit equivalent to such a largest set

Note 1 to entry: Individual languages also encompass *constructed languages* (3.1.10), but do not include formal languages (as defined in ISO 1087:2019, 3.1.10).

Note 2 to entry: Usually, in other contexts, individual languages are simply called “languages”. However, the term “language” has multiple meanings and connotations, which can cause confusion in the context of this document. Still, when an attribute and possibly the plural clearly indicate that individual languages are meant, this document uses only “language(s)”, as in “creole languages”, “Asian languages” or “living languages”.

EXAMPLE English, Guarani, LIBRAS (Língua Brasileira de Sinais/Brazilian Sign Language), Haitian Creole, Esperanto.

3.1.4**individual sign language**

individual language (3.1.3) whose *basic modality* (3.5.11) is the *signed modality* (3.5.4)

Note 1 to entry: Usually “sign language” is part of the name of the respective individual sign language.

EXAMPLE American Sign Language (ASL), langue des signes québécoise/Quebec Sign Language (LSQ).

Note 2 to entry: Individual sign languages differ from the “signed modality” (see 3.5.4), by which an individual language can be expressed which is normally expressed in another *language modality* (3.4.7), such as by “Signing Exact English” for expressing English. Therefore, the term “signed language” is not used as a synonym to the term “individual sign language”.

3.1.5**speaker**

person who is capable of making use of an *idiolect* (3.1.2)

Note 1 to entry: The term “speaker” covers the use of all *language modalities* (3.4.7), and is thus used to denote a generic concept, “speaker”, also covering all specific concepts such as “writer”, “signer”, etc., which can be introduced when needed. The alternative encompassing term “language user”, although technically closer to the intended generic meaning, has proven to render the text much less accessible.

3.1.6**language use event**

event of language use

event in which a *speaker* (3.1.5) expresses themselves by means of an *idiolect* (3.1.2)

Note 1 to entry: Language use events can belong to one of several *language modalities* (3.4.7). A case involving the *spoken modality* (3.5.1) or the *multimodal modality* (3.5.2) is also called a “speech event”. In the case of the *written modality* (see 3.5.3), it is a writing event or an event of producing a written text. In the case of the *signed modality* (3.5.4), it is a signing or signed event, etc.

3.1.7**enhanced communicative functioning ability**

enhanced ability of a *speaker* (3.1.5) during a *language use event* (3.1.6) where the speaker deviates from average communicative functioning by some sort of enhancement

3.1.8**communicative functioning constraint**

constraint of a *speaker* (3.1.5) during a *language use event* (3.1.6) where the speaker deviates from average communicative functioning by being hampered by some limiting factor

Note 1 to entry: A communicative functioning constraint diagnosed as advanced or severe is usually identified as an impairment in the form of a communication disorder affecting the *speaker* (3.1.5).

3.1.9**natural language**

individual language (3.1.3) which is or was in active use in a language community, passed on from one generation of *speakers* (3.1.5) to the next

EXAMPLE Bambara, English, Haitian Creole, Latin, LIBRAS (Língua Brasileira de Sinais/Brazilian Sign Language).

3.1.10**constructed language**

individual language (3.1.3) whose rules are explicitly established prior to its use

EXAMPLE Esperanto, Volapük, Quenya, Na'vi.

Note 1 to entry: Constructed languages do not include reconstructed languages, computer programming languages, mark-up languages or similar formal languages.

Note 2 to entry: Some constructed languages are based on one or several *natural languages* (3.1.9) and are therefore not artificial. Therefore, the term “artificial language”, which is often used as a synonym, is not used in the ISO 21636 series.

3.2 Terms related to linguistic variation and language varieties**3.2.1****linguistic variation**

language variation

differences within and between *individual languages* (3.1.3)

3.2.2**external criterion for linguistic variation**

set of properties of *idiolects* (3.1.2) that are based on factors external to the linguistic features of the idiolects' systems

Note 1 to entry: External criteria for linguistic variation contain properties of idiolects that pertain to the *speakers* (3.1.5) who use the idiolects, or to the *language use event* (3.1.6) in which the idiolects are used.

EXAMPLE “Being characteristic of speakers from East Anglia” is a property which is the only element of an external criterion for linguistic variation [in this case, a criterion related to geographical space (see the example to 3.2.4), defining a certain *dialect* (3.4.1) of English].

3.2.3**structural criterion for linguistic variation**

set of properties of *idiolects* (3.1.2) that are based on the linguistic features of the idiolects' systems

Note 1 to entry: This set of properties includes in particular phonetic, phonological, morphological, syntactic, lexical, semantic or pragmatic properties.

Note 2 to entry: Elements of the structural criterion for linguistic variation are also called “markers”, e.g. in ISO/TR 20694. The term “structural criterion for linguistic variation” is preferred because it integrates better with the framework for *linguistic variation* (3.2.1) developed in the ISO 21636 series.

3.2.4**dimension of linguistic variation**

set of external criteria for *linguistic variation* (3.2.1) of the same kind which can serve to distinguish subsets of *individual languages* (3.1.3)

Note 1 to entry: Criteria are “of the same kind” if they all refer to analogous properties of idiolects (i.e. properties of the same ontological domain) such as, for instance, properties related to a) geographical space, b) time or c) social groups, etc.

Note 2 to entry: The dimensions assumed in the ISO 21636 series framework are listed in 3.3.

EXAMPLE The set of external criteria which all contain properties related to the geographical locations and regions form a dimension of linguistic variation, in this case the *space dimension* (3.3.1), which distinguishes the *dialects* (3.4.1) of individual languages (see the example to 3.2.2).

3.2.5**language variety
variety**

<language coding> largest subset of an *individual language* (3.1.3) that is internally consistent with regard to both an *external criterion for linguistic variation* (3.2.2) and a *structural criterion for linguistic variation* (3.2.3), and that can be identified and named

Note 1 to entry: Since terms such as “linguistic variation”, “language variation”, “linguistic variant”, “language variant” or “linguistic variety” are also used to represent other concepts, only the term “language variety” is used in the ISO 21636 series.

3.3 Terms related to dimensions of linguistic variation**3.3.1****space dimension**

geographical space dimension
dimension of linguistic variation (3.2.4) that refers to geographical locations and regions

Note 1 to entry: *Language varieties* (3.2.5) which can be distinguished according to the space dimension comprise *dialects* (3.4.1) and (supra-regional) *standard varieties* (3.4.2).

3.3.2**time dimension**

dimension of linguistic variation (3.2.4) that refers to spans of time

Note 1 to entry: *Language varieties* (3.2.5) which can be distinguished according to the time dimension are in particular *historical language periods* (3.4.3) and *language epochs* (3.4.4).

3.3.3**social group dimension**

dimension of linguistic variation (3.2.4) that refers to social groups other than geographically defined groups

Note 1 to entry: *Language varieties* (3.2.5) which can be distinguished according to the social group dimension are in particular *sociolects* (3.4.5) and *technolects* (3.4.6). There may be several distinct sociolects or technolects in a given *individual language* (3.1.3).

3.3.4**medium dimension**

dimension of linguistic variation (3.2.4) that refers to the medium used for communication

Note 1 to entry: *Language varieties* (3.2.5) which can be distinguished according to the medium dimension are *language modalities* (3.4.7). Most of them are defined in 3.5.