



**International
Standard**

ISO 21636-3

**Language coding — A framework
for language varieties —**

**Part 3:
Application of the framework**

*Codage des langues — Identification et description des variétés
de langues —*

Partie 3: Exigences et recommandations pour la mise en œuvre

**First edition
2024-06**

STANDARDSISO.COM : Click to view the full PDF of ISO 21636-3:2024

STANDARDSISO.COM : Click to view the full PDF of ISO 21636-3:2024



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, definitions and abbreviated terms	2
3.1 Terms and definitions	2
3.2 Abbreviated terms.....	2
4 Indication of language varieties according to the dimensions of linguistic variation	2
4.1 Overview.....	2
4.2 Indication of individual language varieties.....	3
4.3 Indication of the (geographical) space dimension of linguistic variation.....	4
4.4 Indication of the time dimension of linguistic variation.....	4
4.5 Indication of the social group dimension of linguistic variation.....	5
4.6 Indication of the medium dimension of linguistic variation.....	5
4.7 Indication of the situation dimension of linguistic variation.....	7
4.8 Indication of the person dimension of linguistic variation.....	8
4.9 Indication of the proficiency dimension of linguistic variation.....	8
4.10 Indication of the communicative functioning dimension of linguistic variation.....	8
Bibliography	10

STANDARDSISO.COM : Click to view the full PDF of ISO 21636-3:2024

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.

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^[6].

The metadata categories and values addressed in this document can be candidates for a future fine-grained coding of language varieties based on the comprehensive principles of the ISO 21636 series. Thus, this document 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;

ISO 21636-3:2024(en)

- 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.

STANDARDSISO.COM : Click to view the full PDF of ISO 21636-3:2024

Language coding — A framework for language varieties —

Part 3: Application of the framework

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 gives guidance on how to apply the framework to identify basic dimensions and sub-dimensions of linguistic variation and the resulting varieties, including major modalities of human communication. It does not include any code or individual identifiers.

This document is structured strictly analogously to ISO/TR 21636-2. For a general description of the dimension and varieties dealt with in each clause, the user can refer to the corresponding clause in that document.

This document focuses only on the identification and description of language varieties, not on the general, formal or technical aspects of the description of human language resources (LRs), which are covered by general metadata frameworks.

NOTE 1 For the general description of a language resource, a user can minimally apply at least the metadata of the Open Language Archives Community (OLAC) metadata standard^[2], which provides an application of the Dublin Core metadata element set as defined by the Dublin Core Metadata Initiative (DCMI)^[8]. These descriptors have been recognized in ISO 15836-1:2017.

NOTE 2 The Component Metadata Infrastructure (CMDI)^[9] provides a best practice guide^[10] for the sake of technical and content interoperability between LR as well as of their sustainability.

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 15924, *Information and documentation — Codes for the representation of names of scripts*

ISO 21636-1, *Language coding — A framework for language varieties — Part 1: Vocabulary*

3 Terms, definitions and abbreviated terms

3.1 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 21636-1 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/>

3.2 Abbreviated terms

ACTFL	American Council on the Teaching of Foreign Languages
AAL	ambient assisted living
BCP	Best Current Practice
CEFRL	Common European Framework of Reference for Languages
CMDI	Component Metadata Infrastructure
DCMI	Dublin Core Metadata Initiative
ICT	information and communication technology
IETF	Internet Engineering Task Force
ILR	Interagency Language Roundtable
LR	language resource
LT	language technology
OLAC	Open Language Archives Community

4 Indication of language varieties according to the dimensions of linguistic variation

4.1 Overview

All major metadata sets for the formal description of human communication and LRs include the identification of the individual language(s) of the LR in accordance with ISO 639.

The ISO 21636 series, in turn, specifies which type of descriptors are needed to exhaustively account for the place of an event of language use or an LR in the multidimensional space of linguistic variation within an individual language. Therefore, the ISO 21636 series represents the basis for extensions to the major metadata standards such as the OLAC^[7] or CMDI^[9] metadata formats.

In particular, it can be employed as the basis for a coherent system of extensions to the widely used IETF Best Current Practice 47 (BCP 47)^[11], which specifies “Language Tags”. BCP 47 already includes the identification of language varieties (in particular, dialects, scripts and regions), but the BCP 47 recommendations are far from complete. It is also feasible to define extensions to subtag elements. These extensions may be used to implement the concepts established in this document, for example by establishing one extension for each dimension of linguistic variation, or just one extension that combines both the dimension of linguistic variation and its assigned value.

With the exception of examples and some very general values (e.g. in the medium dimension), this document does not propose any concrete values to be used within the framework defined. A mechanism to register

these values will complement this document in future editions, providing an implementation based on this document.

Certain dimensions of linguistic variation are for all practical purposes open lists, such as dialects (different for each individual language), language periods (potentially different for each individual language) and sociolects (some potentially specific to certain individual languages). According to this document, the language modalities and language registers as well as proficiency levels of learner varieties and communicative functioning abilities and constraints comprise shorter lists. The individual dimension of linguistic variation is already covered by identifying the speaker or speakers, which often happens elsewhere in the metadata and will usually not need to be coded among the language varieties.

It is important for all stakeholders to be aware of the complexity of linguistic variation within individual languages and that at least the eight different dimensions of linguistic variation identified and described in the ISO 21636 series are needed to cope with these complexities. This document develops a conception of individual languages as sets of idiolects, and varieties as subsets that are characterized by external criteria (containing properties related to space, time, social space, medium, etc.) and structural criteria (properties related to pronunciation, lexical items, etc.). This conception is offered as a capable model for dealing with linguistic variation for the purpose of standardization. The need to recognize the establishment of individual varieties is an empirical question, and also practically depends on the application. As such, individual values and resulting tags can be subject to scholarly debate, but this does not take away the need for a general framework of reference.

4.2 Indication of individual language varieties

Any given event of language use, represented in an LR, belongs simultaneously to a certain dialect, to a certain language period, to a certain language modality (it may be written, or oral, etc.), to a certain sociolect, etc. Therefore, the values for each of these dimensions of linguistic variation can and should be stated side by side.

For the sake of optimal re-usability, it is generally advisable to identify language varieties of as many dimensions of linguistic variation as possible used in a given LR, as far as they are known. For this purpose, established conventions and labels for identifying specific varieties should be followed whenever possible.

For the retrieval and re-use of LRs, the specification of language varieties can be crucial, even if they were not the focus of the creators of the LRs at hand. Therefore, the position of a given LR should ideally be made explicit for each of the dimensions of linguistic variation.

However, it is not possible or relevant in all cases to indicate the respective varieties according to all dimensions of linguistic variation. In the course of the description and identification of the language varieties at hand, omission of a dimension of linguistic variation should also be made explicit, for instance by marking them as “unspecified”, or a more specific value, such as “unknown”.

EXAMPLE 1 [dialect:] unspecified.

For the purpose of practicality, it may be established that leaving out a certain dimension of linguistic variation implies that the value (variety) for that dimension of linguistic variation is unspecified, or that it belongs to a certain value – for instance, leaving out the proficiency dimension specification can imply that only native speakers are involved. However, such implicit implications should be made explicit at some prominent place in the general description of the LRs at hand.

The specification of any of the dimensions of linguistic variation can vary with respect to its reliability or certainty status – for example, some values are being assumed or inferred, others are certain or confirmed. This certainty status of the specification should be made explicit by adding a certainty status label immediately to the language variety indication in question, such as “unconfirmed”, or more specifically “assumed” or “inferred”, or similar. This label then indicates a property that refers to the respective statement of the language variety, not itself to the event of language use or the LR in question. In this sense, it is meta-meta-information.

EXAMPLE 2 [dialect:] Bavarian (certainty: inferred).

The fact that an LR contains several events of language use that belong to different language varieties belonging to the same dimension of linguistic variation, such as a dialogue between speakers that use different dialects, should be clearly stated.

Sometimes, the idiolect used in an LR is deliberately chosen to imitate that of different variety than the one that would naturally apply to the speaker or situation. In such cases, the fact that a language variety is applied to imitate another language variety should be indicated by a qualification such as “adopted”, “non-original”, “imitated” or similar. Again, this qualification applies only to the one variety in question and not to the speech-event as such.

EXAMPLE 3 [time period:] Victorian English (imitated).

This document uses explicit unabbreviated labels to refer to the respective varieties used in the examples. Of course, for the purpose of building, for instance, language tags, such detailed identifications can and will be abbreviated. The abbreviated labels should be unique for the specific individual language and across dimensions of linguistic variation to avoid confusion, and they shall be clearly explained at some place, for instance in a registry for such labels once the varieties are centrally registered.

4.3 Indication of the (geographical) space dimension of linguistic variation

The space dimension of an event of language use, or an LR that represents an event of language use, is specified by identifying either the supra-regional standard variety or the dialect characteristic of the geographic region or location where the speaker grew up, and, ideally, the dialect or location where the speaker’s parents grew up. It should be specified as exactly as possible, stating either the name of the dialect or the geographical place or region.

EXAMPLE 1 [dialect:] East Anglian; [subdialect:] Norfolk.

EXAMPLE 2 [dialect:] East Anglian; [dialect of mother:] Wales.

If there is an established name for the language variety of a region, that name should be indicated. For the sake of international communication, in an English document, the English name of the dialect should be indicated, possibly in addition to the original name of the dialect in the language itself, or in another relevant individual language. Analogous solutions hold true for documents in other languages.

When needed, a certainty status label should be added. In particular, if the specification of the dialect lacks confirmation, this uncertainty of the specification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”, or similar. Similarly, the label “imitated” can be added (see ISO/TR 21636-2:2023, 4.2).

EXAMPLE 3 [dialect:] Irish English; [subdialect:] unidentified.

EXAMPLE 4 [dialect:] Irish English; [subdialect:] Dublin English; [status of subdialect:] assumed.

Another option, in particular if there is no established name for the dialect, is to state the geographical region and place that is most characteristic of the speaker of the event of language use in a given LR. This holds true, in particular, for varieties used by expatriates or migrants (non-traditional dialects) which are characterized by influence from other individual languages. These may also be explicitly stated.

EXAMPLE 5 [language:] Urdu; [dialect:] London; [influence by language:] English.

4.4 Indication of the time dimension of linguistic variation

The time dimension of linguistic variation should be specified as exactly as possible. If there is an established name for the language period or language epoch of an event of language use, that name should be indicated. It is possible to make several indications, starting with the larger language epoch and becoming more specific. In principle, the most specific value implies the others, but these can still be useful for other purposes. Also, in periods of transition, it can be useful to state whether a given event or resource can be more clearly grouped into one language period than another.

For the sake of international communication, in an English document, the English name of the given language period or language epoch should be indicated (in addition to the original name of the language period

or language epoch in the native individual language itself, or in another individual language). Analogous solutions hold true for documents in other languages.

EXAMPLE 1 [epoch:] modern English; [period:] Victorian English.

When needed, a certainty status label should be added. In particular, if the specification of the language period or language epoch lacks confirmation, this uncertainty of the specification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”. Similarly, the label “imitated” can be added (see ISO/TR 21636-2:2023, 4.2).

EXAMPLE 2 [epoch:] modern English; [period:] early 20th century.

EXAMPLE 3 [epoch:] modern English; [period:] early 20th century; [status of period:] assumed.

NOTE The date and the time when an LR was created, for instance by digitizing the recording of a communication event, can differ from the date of the event of language use itself, but both dates can be recorded in the metadata for the LR. The date of the original language use event applies for the identification of the relevant temporal variety. In some cases, several dates have to be distinguished – for instance in the case of a digitization of a 19th century re-edition of a 17th-century text. In such a case, the language period to be indicated according to this framework would be the 17th century, which is when the original event of language use, the writing of the text, happened.

When appropriate, date forms specified in the ISO 8601 series should be used.

4.5 Indication of the social group dimension of linguistic variation

The sociolect(s) used in an event of language use or in an LR should be specified as exactly as possible. In general, the social stratum or subcommunity in which the speaker was socialized should be stated as precisely as possible. Ideally, reference to the speaker’s parents’ social groups should be included as well. If there is an established name for the sociolect of a speaker in an event of language use, that name should be used. Depending on the topic of the event of language use or the LR, it may be most relevant to state the education or occupational group of the speaker or speakers.

For the sake of international communication, in an English document, the English name of the sociolect should be indicated (in addition to the original name of the sociolect in the native individual language itself or in another individual language). Analogous solutions hold true for documents in other languages.

EXAMPLE 1 [sociolect:] African American vernacular English.

EXAMPLE 2 [sociolect:] middle class English; [sociolect/education:] academic.

EXAMPLE 3 [sociolect/technolect:] (English) business speak.

There is a close interplay between sociolects and registers – some situations involve certain social groups, such as corporate jargon or “legalese”, which are bound to the respective occupational groups of businesspersons or lawyers and the like (sociolect), where the sociolects or technolects most typically occur when these are among themselves (register). Still, the dimensions of linguistic variation should be kept apart as far as possible: the sociolect is determined more by the general background than by the specific situation, which is covered by a register. For instance, the Japanese spoken by the Japanese royal family was historically different from that of most other Japanese speakers (sociolect), but it still differed according to the situation. For instance, the higher register used during a formal ritual was different from the lower register used in an informal chat within the Japanese royal family.

When needed, a certainty status label should be added. In particular, if the specification of the sociolect lacks confirmation, this uncertainty of the specification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”. Similarly, the label “imitated” can be added (see ISO/TR 21636-2:2023, 4.2).

EXAMPLE 4 [sociolect:] middle class English; [status of sociolect:] inferred.

4.6 Indication of the medium dimension of linguistic variation

The language modality should be specified as exactly as possible. If there is an established name for the language modality of an event of language use, that name should be indicated.

For the sake of international communication, in an English document, the English name of the language modality should be used (in addition to the original name of the language modality in the individual language itself or in another individual language). Analogous solutions hold true for documents in other languages.

EXAMPLE 1 [modality:] spoken.

For this dimension of linguistic variation, several possible values are defined in ISO 21636-1:2024, 3.5, but more can be added, for example based on the instrument used to communicate acoustically by so-called surrogate (language) modalities of the spoken modality.

EXAMPLE 2 [modality:] whistling; [tool:] grass blade.

EXAMPLE 3 [modality:] instrumental; [instrument:] bells.

The label “spoken [modality]” as in Example 1 should only be used if the speech event did not include any visual communication, e.g. when speaking on the phone, recording audio messages, or speaking across visual obstacles or larger distances. For regular oral communication with visual contact, the label “multimodal [modality]” should be used, even if the LR only captures the audio signal.

If necessary, subtypes of language modalities can be indicated. This holds in particular for the spoken and written modalities. In both cases, there is a certain overlap with text genres (see below in this subclause and [4.7](#)).

In the case of spoken modalities, it should be possible to indicate different subtypes, some of which also interact with the communicative functioning dimension, which is generally reserved for involuntary phenomena. Some general subtypes are yelling, whispering and in particular singing, which in turn can also have a number of sub-types (see Example 6).

EXAMPLE 4 [modality:] multimodal; [mode:] whispering.

EXAMPLE 5 [modality:] spoken; [medium:] audio-chat-message.

EXAMPLE 6 [modality:] multimodal; [mode:] singing; [modal type:] singing with tones preserved.

Even more differentiation can be necessary for the written modality, especially in recent times with the many digital (computer-aided) forms of communication, either one-way (e.g. websites, blog-posts) or two-way (text-messages/chat, comments on social media platforms, etc.).

EXAMPLE 7 [modality:] written; [medium:] handwritten.

EXAMPLE 8 [modality:] written; [medium:] typed/digital; [form:] chat message.

EXAMPLE 9 [modality:] written; [medium:] typed/digital; [form:] blog post.

The examples above distinguish among:

- a) the more general language modalities and the more specific modes, focusing on choices of the speaker (see Examples 4 and 6);
- b) the more specific medium indications, focusing on the pure physical means (see Examples 5, 7, 8 and 9);
- c) the more specific form, focusing on the type of communication and platform (see Examples 8 and 9).

Especially regarding form, there is some overlap with text genres; for example, messages in chats or other social media are characterized by both a specific medium (medium dimension) and a text genre (situation dimension). Therefore, when tagging a given collection of LRs, these indications should be given consistently with respect to one or the other dimension, depending on whether the medium aspect or the genre or text type aspect is more in focus. In some cases, it can be necessary to make specifications concerning both dimensions of linguistic variation.

As is the case of other choices in the examples above, this indicates possible and suggested consistent ways of how to indicate the relevant information, but does not exclude the implementation of other consistent systems within this dimension of linguistic variation. Before offering a finalized and possibly complete system for specifying the medium dimension, the categories and subcategories as well as their possible values shall be worked out carefully.

If the specification of the language modality lacks confirmation, this uncertainty of the specification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”. Similarly, the label “imitated” can be added (see ISO/TR 21636-2:2023, 4.2).

EXAMPLE 10 [modality:] multimodal; [status of modality:] inferred.

Example 10 applies for instance in the case of an audio recording where it can be inferred that the speakers were able to see each other while they were speaking, which can be relevant.

In cases where more than one language modality represented in a given LR, it can be useful to state all modalities (if possible), and to indicate the primary and secondary language modalities. The primary language modality is the one in which the linguistic expression was first created, the secondary language modality is the one in which it was re-created or cited.

Example 11 holds in the case of a written text read aloud.

EXAMPLE 11 [modality/primary:] written; [modality/secondary:] multimodal.

Example 12 holds in the case of a transcript of a dialogue.

EXAMPLE 12 [modality/primary:] spoken; [modality/secondary:] written.

Sometimes written communication events are transcribed into a different writing system than the original writing system of the individual language or variety. Frequently utterances in non-written languages or varieties are transcribed. For this transcription, different writing systems and different transcription conventions may be applied. For the identification of regular writing systems, ISO 15924 shall apply.

When more than one transcription convention is employed, the name of each given transcription system should be indicated. For the sake of international communication, in English documents, the English name of the transcription system of a given language modality should be indicated, in addition to the original name of the transcription. BCP 47^[11] gives advice on how to identify transcription systems, which change from language period to language period (see 4.4 about the time dimension).

When needed, a certainty status label should be added. In particular, if the specification of the medium dimension of a language variety – or of its transcription systems – lacks confirmation, this uncertainty of the specification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”. This will usually only refer to a primary modality, for instance, when it is only inferred that the LR contains a reading aloud of an originally written text.

The imitation label does not apply to the medium dimension.

EXAMPLE 13 [modality/primary:] written; [status of primary modality:] inferred; [modality/secondary:] spoken.

4.7 Indication of the situation dimension of linguistic variation

The language register (degree of formality, type of situation) should be specified as exactly as possible. If there is an established name for the language register of an event of language use, that name should be used.

For the sake of international communication, in an English document, the English name of the register should be indicated (in addition to the original name of the register in the individual language itself or in another individual language). Analogous solutions hold true for documents in other languages.

EXAMPLE 1 [register:] informal; [register/context:] familiar.

EXAMPLE 2 [register:] neutral; [register/addressee:] child-directed speech.

EXAMPLE 3 [register:] neutral; [register/accessibility:] plain language.

If not covered elsewhere in the metadata, genres should be indicated as part of the space dimension. There is a certain overlap with the medium dimension (see 4.6); singing and computer-aided forms of communication in particular can require indications under both dimensions.

EXAMPLE 4 [genre:] lyrics; [situation:] singing.

When needed, a certainty status label should be added. In particular, if the specification of the register lacks confirmation, this uncertainty of the specification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”. Similarly, the label “imitated” can be added (see ISO/TR 21636-2:2023, 4.2).

EXAMPLE 5 [register:] formal; [status of register:] inferred.

4.8 Indication of the person dimension of linguistic variation

The personal variety should be specified by identifying the speaker, if known.

The identification of the speaker is usually covered by the general metadata for the respective LRs. If this is not the case, it can be done as part of the framework of this document.

EXAMPLE 1 [person:] John Doe.

When needed, a certainty status label should be added. In particular, if the specification of the person lacks confirmation, this uncertainty of the specification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”. Similarly, the label “imitated” can be added (see ISO/TR 21636-2:2023, 4.2).

EXAMPLE 2 [person:] John Doe; [status of person:] assumed.

4.9 Indication of the proficiency dimension of linguistic variation

The learner varieties or native proficiency variety can be indicated by specifying a proficiency profile (i.e. a combination of a type of language acquisition process and a stage of learning). These values should be specified as precisely as possible. “Language acquisition process” refers to first- (L1) or second- (L2) language acquisition, or even foreign language acquisition, if this is differentiated from L2-acquisition. If there is an established name for the proficiency profile of a speaker at the time of an event of language use, that name should be indicated.

For the sake of international communication, in an English document, the English name of the proficiency profile should be used (in addition to the original name of the proficiency profile in the individual language itself or in another individual language). Analogous solutions hold true for documents in other languages.

For some individual languages, the proficiency profiles are regulated by normative systems such as ILR^[12], ACTFL^[13], Canadian Language Benchmarks or CEFRL^[14], among many others. If such a system is applied, it should be named together with the name of the proficiency profile.

EXAMPLE 1 (no specific framework, impressionistic labelling): [acquisition:] second language acquisition; [proficiency stage:] beginner.

EXAMPLE 2 [acquisition:] second language acquisition; [proficiency stage:] elementary / A2 (CEFRL).

EXAMPLE 3 (underlying theory not indicated): [acquisition:] first language acquisition; [proficiency stage:] two-word-stage.

When needed, a certainty status label should be added. In particular, if the specification of the stage of language learning lacks confirmation, this uncertainty of the specification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”. Similarly, the label “imitated” can be added (see ISO/TR 21636-2:2023, 4.2).

EXAMPLE 4 [acquisition:] second language acquisition; [acquisition/context:] school; [proficiency stage:] intermediate; [status of proficiency stage:] inferred.

Most often, the proficiency profile will not be mentioned explicitly, implying instances of language use by a fluent native speaker. Ideally, such an implication is made explicit somewhere.

4.10 Indication of the communicative functioning dimension of linguistic variation

The value on the communicative functioning dimension (enhanced abilities or constraints) should be specified as exactly as possible. If there is an established name for an enhanced communicative functioning