FINAL
DRAFT

TECHNICAL
REPORT

ISO/DTR 21636-2

ISO/TC **37**/SC **2**

Secretariat: **SCC**

Voting begins on:
2022-12-14

Voting terminates on:
2023-02-08

Language coding — A framework for language varieties —

Part 2:
Description of the framework

Reference number
ISO/DTR 21636-2:2022(E)

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

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, 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 is complementary to 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 (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 one from another. In addition, linguistic variation also exists within each individual language. No language use event is independent of this language-internal linguistic variation. This variation is also present in LRs. Therefore, 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 regarding to its location in each of these dimensions of linguistic variation.

Some language varieties can again be differentiated into sub-varieties of the same or other dimensions of linguistic variation. This level of detail is usually not the focus of this document.

The borders between individual languages are sometimes hard to establish. In particular, different groups of people can disagree with regard to their assessment of a given set of idiolects as forming an individual language or not, depending on their assessment of the mutual intelligibility and/or of the socio-political situation. Therefore, there are cases where the status of a given set of idiolects as constituting an individual language, or a language variety, or a group of closely related individual 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 language-internal linguistic variation, that is, of language varieties and sub-varieties. It does not aim at establishing all language varieties of all individual languages, which at any given point in time would amount to an immense list. But it aims at creating an exhausting list of the types of descriptors that are needed for precisely and completely characterizing the status regarding linguistic variation of a given language use event, for instance an utterance or an LR. It does so by describing all the dimensions of linguistic variation in which individual languages can vary internally, and by indicating the major resulting language varieties that typically occur in individual languages.

Technically, individual languages are conceived in this document as sets of idiolects, which in turn are sets of expressions of human language, each expression with its meaning, where each idiolect is characterized by a coherent system of structural features. This understanding of “idiolect” avoids difficulties of other approaches, particularly those where the term “idiolect” is used in the sense of what in this document is named a “personal variety”: the total repertoire a speaker is able to use in a language. However, a personal variety (in such approaches called “idiolect”) usually is internally heterogeneous (it varies in particular according to different situations and/or media) and therefore not suitable to serve as the smallest unit of linguistic variation. In contrast, idiolects according to the framework developed in this document are homogeneous by definition, whereas personal varieties are sets of idiolects.

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 few major types and constitute the different dimensions of linguistic variation; there is one dimension of linguistic variation for each relevant major type of external properties of idiolects (such as properties referring to geographical space, properties referring to time, etc.).

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[17].

4.2 Dimensions of linguistic variation

Linguistic manifestations such as LRs and the events of language use themselves can be characterized according to the following dimensions of linguistic variation (see ISO 21636-1:—, 3.3):

a) space dimension;

b) time dimension;

c) social group dimension;

d) medium dimension;

e) situation dimension;

f) individual speaker dimension;

g) proficiency dimension;

h) communicative functioning dimension.

This document characterizes each of the dimensions and their respective varieties in a general manner; instructions on how to identify and indicate the respective varieties belonging to each of these dimensions are given in ISO 21636-3. The structure of this document and ISO 21636-3 is strictly parallel so that the reader can easily compare the general discussion and the instructions.

Although there can be interferences between some dimensions, the dimensions of linguistic variation are in principle independent of one another. A complete characterization of any given language use event and any LR with regard to linguistic variation would identify the respective varieties to which the event or resource belongs in all and each of these eight dimensions of linguistic variation. In practice, a description can focus only on a few salient or relevant dimensions of linguistic variation, and (perhaps tacitly) assume default values for other dimensions of linguistic variation.

For reasons of readability, “speaker” is always used generically in this document, covering also more specific concepts such as “writer”, “signer”, etc. Equally, other comments referring to properties specific to speaking always hold analogously for the other language modalities.

Language varieties of some dimensions of linguistic variation can have sub-varieties, usually of the same dimension, such that a given language use event or a given LR can have more than one value with respect to a dimension of linguistic variation. Where two varieties of the same dimension apply, usually one variety is more specific and the other broader. For instance, a recording of a speaker using a Norfolk dialect can also be characterized as belonging to the broader East Anglian dialect of English.

There are even examples for sub-varieties of different dimensions. For instance, the middle Bavarian period of the Bavarian dialect of German is different from the Bavarian dialect combined with the middle German period. Such cases are not further addressed in this document.

There can also be mixed events of language use or LRs that contain several events of language use that belong to different varieties according to the same dimension of linguistic variation. For instance, a dialogue between speakers uses different dialects, or a dictionary covers several dialects, sociolects, etc. of an individual language. In such cases, all respective (groups of) language varieties will have to be identified, and if possible, the respective parts of such a LR will have to be related to their respective language varieties; for instance, according to the different participants and to the different time segments of such a dialogue, or to the different entries in such a dictionary.

Another special case involves LRs where the language use is "non-native" in the sense that a speaker is deliberately imitating a language variety different from his/her own variety. For instance, a speaker of a dialect X imitates speech of another dialect Y, or a speaker imitates another speaker, or an author makes use, for example as a stylistic device, of linguistic expressions as conceived as typical for a certain historical language period, or of a certain social group to which the writer does not belong. In such cases, it is advisable to state the fact that the variety in question is being imitated. (In general, such cases can be dealt with in a way compatible with Extension T of IETF BCP 47[11].)

4.3 The space dimension and its varieties

With respect to the space dimension, an individual language can be differentiated into dialects (and these into sub-dialects) and sometimes also a supra-regional standard variety. This is often the most complex and differentiated kind of linguistic variation.

EXAMPLE 1 (English dialect): East Anglian.

EXAMPLE 2 (supra-regional standard variety): United Kingdom Standard English.

When determining the dialect of a speaker, the factor that is mainly considered is the geographical region of the socialization of the speaker, that is, where the speaker grew up and also where the speaker's parents grew up. If these factors are heterogeneous, for example due to migration or due to parents from different regions, sociolinguists try to identify the major dialect which the language use of the speaker most strongly resembles. In such cases, also minor dialectal influences can be indicated, for instance from the region of (one of) the speaker's parents, or of a region where the speaker moved to in a later phase of his or her life. The geographical region of an individual language can cover several countries and continents, giving rise to new dialects.

EXAMPLE 3 Western American English with influence from southern British English.

The names given to individual dialects are often traditional and usually refer to the geographical region where the dialect is spoken. When appropriate, the ISO 3166 series can be applied for countries and established regions, or ISO 19111 and ISO 19112 for spatial referencing. How many and which dialects are to be distinguished on a given level of specificity is often debated between specialists, and so are the names and the borders of the dialects.

Most major and also many smaller, well-researched individual languages are characterized by a finite number of established major dialects. In the case of little known or disputed language varieties, the identification of a dialect can be supported by a reference to some scholarly work where the dialect is established or identified.

EXAMPLE 4 Cheshire English [cf. Leigh E. Introduction in: A Glossary of Words Used in the Dialect of Cheshire (Hamilton, Adams, and Co./Minshull and Hughes; 1877)].

Dialect areas can overlap and often can best be defined by similarity with some prototypical core variety. Hence, belonging to a certain dialect can be a question of degree for an idiolect and does not necessarily exclude it from also being acceptable as a member of another dialect to a certain degree.

In individual languages used in a larger geographical area, there is often one variety, usually based on one specific traditional dialect or a group of dialects, that is recognized as "standard" by most or

all speakers across the whole or a larger part of the geographical area of the individual language. The standard variety is often characterized by a high degree of normalization and is used in official communication. In such cases, many speakers can use both a local dialect and the standard variety. Again, in the case of the standard variety, often the influence of a local dialect is still evident (for instance as an "accent"), even with speakers who do not have strong mastery in the original local dialect.

EXAMPLE 5 High German with a Bavarian accent.

In the case of individual languages with a very broad geographical coverage, especially those used in different countries (e.g. German) and continents (in particular Arabic, Chinese, Dutch, English, French, Portuguese and Spanish) there can be more than one standard variety.

A special case is posed by diaspora varieties spoken in a geographical region where other individual languages are more strongly present. Usually the influence of the dominant language is evident in the individual language in such a situation. This again can influence the native individual language of speakers who live for a while in such an area, even when they speak their original dialect.

EXAMPLE 6 Urdu spoken in London (with influence from English).

When speakers speak several dialects and/or the standard variety more or less fluently, it can be necessary to determine which of the dialects or standard variety has been used in each language use event. For some purposes, it can additionally be useful to indicate which other individual languages and which other dialects of the individual language in question the speakers are able to speak, because this can influence their linguistic behaviour, even when speaking another dialect or a standard variety.

4.4 The time dimension and its varieties

With respect to the time dimension, an individual language can be differentiated into language epochs and historical language periods. They can be named after the eras of political organization, of rulers, or of cultural, social or economic development. Language epochs can comprise distinct language periods.

For several well-studied individual languages, in particular individual languages with a long tradition of writing, scholars have established the language epochs "old X", "middle X" and "modern X", where X stands for the name of the individual language. These language epochs are then often sub-divided into "early" and "late" language periods.

EXAMPLE 1 (epoch): Early Middle English.

EXAMPLE 2 (period): Victorian English.

Language epochs can vary in their temporal extension between individual languages. Sometimes they vary even between different dialects within one individual language. This holds even more so for the historical language periods.

The establishment of historical language periods can vary between different experts and depends on their interest or purpose. The beginning and end of a language period are usually not exact points, so that a period can be characterized by vague delineations or prototypes, for example: "the period around the 1880s" or "the 16th and early 17th century". The closer to the present moment, the shorter are the periods of an individual language or language variety that can be distinguished due to more detailed knowledge of the structural features of the individual language. Still, language periods typically span some decades up to a few centuries.

4.5 The social group dimension and its varieties

With respect to the social group dimension, an individual language can be differentiated into sociolects. Sociolects refer to the socialization of speakers as belonging to a certain social group, such as class, milieu, professional group, age group, religious group, ethnic group (if not accounted for by dialects) or gender.

The number and specificity of sociolects that need to be distinguished varies very much from individual language to individual language and reflects the social structure and in particular the social segregation