Ch. de Blandonnet 8, CP 401, 1214 Vernier, Geneva, Switzerland | T: +41 22 749 01 11 | iso.org | central@iso.org

Form 4: New Work Item Proposal

Circulation date: 6-05-2016 Closing date for voting: 06-08-2016	Reference number: ISO XXXXX. (to be given by Central Secretariat)
Proposer TC 37/SC 2	**ISO/TC 37/SC** 2 ☐ Proposal for a new PC
Secretariat TC 37/SC 2	

A proposal for a new work item within the scope of an existing committee shall be submitted to the secretariat of that committee with a copy to the Central Secretariat and, in the case of a subcommittee, a copy to the secretariat of the parent technical committee. Proposals not within the scope of an existing committee shall be submitted to the secretariat of the ISO Technical Management Board.

The proposer of a new work item may be a member body of ISO, the secretariat itself, another technical committee or subcommittee, an organization in liaison, the Technical Management Board or one of the advisory groups, or the Secretary-General.

The proposal will be circulated to the P-members of the technical committee or subcommittee for voting, and to the O-members for information.

IMPORTANT NOTE: Proposals without adequate justification risk rejection or referral to originator.
Guidelines for proposing and justifying a new work item are contained **in Annex C of the ISO/IEC Directives, Part 1.**

☐ The proposer has considered the guidance given in the Annex C during the preparation of the NWIP.

Proposal (to be completed by the proposer)

Title of the proposed deliverable.

English title: Identification and description of language varieties

French title (if available): Identification et description des variétés de langue

Scope of the proposed deliverable.

This proposed standard gives the general principles for the identification and description of varieties of natural language or other human communication means fully or largely equivalent to natural language. It also provides a system of basic dimensions (and sub-dimensions for some of these) as well as core values necessary to specify these dimensions or sub-dimensions. This allows for a systematic and comprehensive identification and description of the manifold kinds of language varieties – down to the language variety of an individual speaker.

Purpose and justification of the proposal*

Description of language resources (LRs – in the broadest sense include written, spoken, signed, and other modalities), is a domain which is gaining importance in connection with the preservation of the digital heritage. More and more LRs – including also new forms of language resources such as social media data – are made available on the Internet. Existing LRs are increasingly digitized, while many new resources are created in digital form from the outset.
In addition, clearly described and identified LRs are a prerequisite for further developing speech technologies to be used use by ever more and smaller language communities, as well as by those suffering from one or the other kind of speech anomaly or communication disorder.

Speech technology is an important part of the language industry which is one of the fastest growing application areas of the ICT industry. The clear metadata approach of the proposed standard is a prerequisite for the sustainability of the archiving of LRs as well as of the interoperability (i.e. re-usability and re-purposability) of LRs over time and with different technologies for different purposes.

A standardized set of dimensions - as the one provided by the proposed standard - for the identification of language varieties is important to guarantee frictionless exchange of information as well as to indicate the degree of re-usability and re-purposability of LRs. In connection with these dimensions LRs are applicable in eBusiness, eHealth, eGovernment, eInclusion, eLearning, smart environments, ambient assisted living (AAL), and virtually all other information-rich applications where LRs are involved.

Consider the following: Is there a verified market need for the proposal? What problem does this standard solve? What value will the document bring to end-users? See Annex C of the ISO/IEC Directives part 1 for more information.

See the following guidance on justification statements on ISO Connect:
https://connect.iso.org/pages/viewpage.action?pageId=27590861

Preparatory work (at a minimum an outline should be included with the proposal)

X A draft is attached ☐ An outline is attached ☐ An existing document to serve as initial basis

The proposer or the proposer's organization is prepared to undertake the preparatory work required:

X Yes ☐ No

If a draft is attached to this proposal,:

Please select from one of the following options (note that if no option is selected, the default will be the first option):

☐ Draft document will be registered as new project in the committee's work programme (stage 20.00)

X Draft document can be registered as a Working Draft (WD – stage 20.20)

☐ Draft document can be registered as a Committee Draft (CD – stage 30.00)

☐ Draft document can be registered as a Draft International Standard (DIS – stage 40.00)

Is this a Management Systems Standard (MSS)?

☐ Yes X No

NOTE: if Yes, the NWIP along with the Justification study (see Annex SL of the Consolidated ISO Supplement) must be sent to the MSS Task Force secretariat (tmb@iso.org) for approval before the NWIP ballot can be launched.

Indication(s) of the preferred type or types of deliverable(s) to be produced under the proposal.

X International Standard ☐ Technical Specification

☐ Publicly Available Specification ☐ Technical Report

Proposed development track

☐ 1 (24 months) X 2 (36 months - default) ☐ 3 (48 months)

Note: Good project management is essential to meeting deadlines. A committee may be granted only one extension of up to 9 months for the total project duration (to be approved by the ISO/TMB).

Known patented items (see ISO/IEC Directives, Part 1 for important guidance)

☐ Yes X No

If "Yes", provide full information as annex

Co-ordination of work: To the best of your knowledge, has this or a similar proposal been submitted to another standards development organization?

☐ Yes X No

If “Yes”, please specify which one(s):

Click here to enter text.

A statement from the proposer as to how the proposed work may relate to or impact on existing work, especially existing ISO and IEC deliverables. The proposer should explain how the work differs from apparently similar work, or explain how duplication and conflict will be minimized.

The proposed standard is complementary to the multipart international standard ISO 639 which focuses on languages, language families and language groups. Based on best practice in the field of documentation of language resources, it provides a generic framework for identifying and describing the language varieties falling under any given language down to the language use of individuals.

It adds a higher level of granularity to the International Standard ISO 24622-1:2015 developed by ISO/TC 37/SC 4.

It is complementary to the work of ISO/TC 37/SC 1/WG 4.

The project leader will work closely with the TC37 Terminology Coordination Group in identifying and defining key terms relevant to language varieties.

A listing of relevant existing documents at the international, regional and national levels.

ISO 639 parts 1-5

ISO/TS 24620-1:2015

ISO 15924:2004

ISO 24622-1:2015

BCP 47

A simple and concise statement identifying and describing relevant affected stakeholder categories (including small and medium sized enterprises) and how they will each benefit from or be impacted by the proposed deliverable(s)

As the proposed standard covers systematically and comprehensively all potential language varieties - including certain kinds of non-verbal communication - it is important for the development of the user interfaces of ICT devices, web design, speech technology, media design, etc.

It is of particular importance in terms of general methodology and system design for language documentation and archives, as well as specialized libraries and academic research in the field of human communication resources.

The proposed standard also complies with the requirements of eAccessibility and eInclusion, as legally required from all signatories of (i.e. states signing and/or ratifying) the UN Convention on the Rights of Persons with Disabilities (CRPD).

Liaisons:

A listing of relevant external international organizations or internal parties (other ISO and/or IEC committees) to be engaged as liaisons in the development of the deliverable(s).

ISO TC 37/SC 2/WG 1

ISO/TC 37/SC 1/WG 4

ISO TC 37/SC 4/WG 1

ISO/TC 37/TCG

Joint/parallel work:

Possible joint/parallel work with:

☐ IEC (please specify committee ID)

Click here to enter text.

☐ CEN (please specify committee ID)

Click here to enter text.

☐ Other (please specify)

Click here to enter text.

A listing of relevant countries which are not already P-members of the committee.

Click here to enter text.

Note: The committee secretary shall distribute this NWIP to the countries listed above to see if they wish to participate in this work

Proposed Project Leader (name and e-mail address)

Sebastian Drude

sebastian.drude@clarin.eu

Name of the Proposer

ISO TC37 SC2 Secretariat & Infoterm

Maryse Benhoff

mmb@bgcommunications.ca

Christian Galinski

christian.galinski@chello.at

This proposal will be developed by:

X An existing Working Group (please specify which one): ISO/TC 37/SC 2/WG 1

☐ A new Working Group (title: Click here to enter text.)

(Note: establishment of a new WG must be approved by committee resolution)

☐ The TC/SC directly

☐ To be determined

Supplementary information relating to the proposal

X This proposal relates to a new ISO document;

☐ This proposal relates to the adoption as an active project of an item currently registered as a Preliminary Work Item;

☐ This proposal relates to the re-establishment of a cancelled project as an active project.

Other:

Click here to enter text.

X Annex(es) are included with this proposal (give details)

WD.

2016-03-29 SD+GA, v1-2_GA

Identification and description of language varieties

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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

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.

ISO XXX was prepared by Technical Committee ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 2, *Terminographical and lexicographical working methods*. It is complementary to the multipart standard ISO 639 Codes for the representation of names of languages.

Table of contents

Introduction

Detailed and exact characterization of linguistic varieties used in a given event of language use, is a domain which is gaining importance, in particular in connection with more and more language resources (LR, in the broad sense of human language which includes written, spoken, and signed language as well as language in other modalities – including also new forms of language use, such as in social media and similar forms of digital communication). Existing language resources are increasingly digitized, while many new resources are created in digital form from the outset.

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

- exchange of secondary information (i.e. bibliographic description data) for making the information on existing LR widely available in a harmonized form,
- researchers are looking for the primary data (i.e. the resources themselves) for many different research purposes, in particular research on linguistic variation,
- language technologies (LT), in particular speech recognition and language analysis and their equivalents in other modalities, are entering more and more dimensions of human communication and need LR for the development of new language technologies and for testing purposes.

For the above-mentioned and further goals and purposes, a standardized set of metadata for the identification of language varieties is important to guarantee frictionless exchange of secondary information as well as to indicate the degree of re-usability and re-purposability of language resources, and the applicability of language technology to a given situation or language resource. These metadata are applicable in eBusiness, eHealth, eGovernment, eInclusion, eLearning, smart environments, ambient assisted living (AAL), and virtually all other information-rich applications where HCRs are involved. A clear metadata approach is also a prerequisite for the sustainability of the archiving of language resources (in particular in the case of cultural heritage and scientific research data) as well as of the interoperability (i.e. re-usability and re-purposability) of LRs and LT over time and with different technologies for different purposes.

This standard is complementary to the ISO 639 series of standards by extending the framework available in order to allow for the identification of language varieties (including geographical, social etc. varieties). The identification of language varieties can then be included in general / library / archival metadata for describing LRs (which may also include technical modalities, time and location of recording, etc., which are not part of this standard).

The provisions of this standard cover

- general rules for the identification and description of language varieties,
- a set of dimensions – some of them having sub-dimensions – and open-ended or closed lists of values that can be assigned to the respective dimension or sub-dimension,
- a set of metadata categories and examples for the respective possible values grouped according to the most important aspects of the description of instances of language use and resulting language resources, related to linguistic variation.

The metadata categories and values addressed in this document may be candidates for a future high-granular coding of language varieties based on comprehensive principles. Thus, this International Standard complies to the “Recommendation on software and content development principles 2010”. (see Annex A)

Stakeholders include, but are not limited to:
- ICT industry (incl. language technologies)
- libraries
- media industry (incl. entertainment)
- WWW communities
- language documentation & archives
- linguistic research
- language training
- emerging new user communities

1 Scope

ISO XXX gives the general principles for the identification and description of varieties of natural language or communication means fully or largely equivalent to human language in the sense of flexible, re-combinable means of communication able to describe complex situations and events and capable to express complex thoughts. It, therefore, excludes
- artificial means of communication with or between machines such as programming languages; and
- those means of human communication which are not fully or not largely equivalent to human language such as individual symbols or gestures that carry isolated meanings but cannot be freely combined into complex expressions.

ISO XXX provides basic dimensions and sub-dimensions for some of these for the identification and description of language varieties as well as core values necessary to specify these dimensions or sub-dimensions.

2 Normative references

Normative references include the following.
For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 639-1:2002, Codes for the representation of names of languages — Part 1: Alpha-2 code
ISO 639-2:1998, Codes for the representation of names of languages — Part 2: Alpha-3 code
ISO 639-3:2007, Codes for the representation of names of languages — Part 3: Alpha-3 code for comprehensive coverage of languages
ISO 639-4
ISO 639-5:2008, Codes for the representation of names of languages — Part 5: Alpha-3 code for language families and groups

ISO 3166-1:2006, Codes for the representation of names of countries and their subdivisions — Part 1: Country codes
ISO 3166-2:2007, Codes for the representation of names of countries and their subdivisions — Part 2: Country subdivision code
ISO 3166-3:1999, Codes for the representation of names of countries and their subdivisions — Part 3: Code for formerly used names of countries
ITU-T X.1081:2004, The telebiometric multimodal model — A framework for the specification of security and safety aspects of telebiometrics
IEC 80000-14:2008, Quantities and units — Part 14: Telebiometrics related to human physiology
ISO 8601:2004, Data elements and interchange formats — Information interchange — Representation of dates and times
ISO/IEC 11179-2:2005, Information technology — Metadata registries (MDR) — Part 2: Classification
ISO/IEC 11179-3:2003, Information technology — Metadata registries (MDR) — Part 3: Registry metamodel and basic attributes
ISO/IEC 11179-4:2004, Information technology — Metadata registries (MDR) — Part 4: Formulation of data definitions
ISO/IEC 11179-5:2005, Information technology — Metadata registries (MDR) — Part 5: Naming and identification principles
ISO/IEC 11179-6:2005, Information technology — Metadata registries (MDR) — Part 6: Registration ISO 12620:2009, Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources
ISO 15836:2009, Information and documentation — The Dublin Core metadata element set
ISO 15924:2004, Information and documentation — Codes for the representation of names of scripts
ISO 19111:2007, Geographic information — Spatial referencing by coordinates
ISO 19112:2003, Geographic information — Spatial referencing by geographic identifiers
ISO 24619:2011, Language resource management — Persistent identification and sustainable access (PISA)
IETF BCP 47 Tags for Identifying Languages
https://en.wikipedia.org/wiki/Real-time_locating_system

3. Terms and definitions

3.1 Language and languages

natural language
language (3.xxx) for human communication that is not an **artificial language** (3.xxx)
[SOURCE: ISO/FDIS 639-4:2009, 3.27 – modified by adapting the ordering number according to Chapter 3 of this standard]

human language
systematic use of sounds, characters, symbols or signs to express or communicate meaning or a message between humans

NOTE 1 to entry: This definition is intended to serve as a working definition for the purpose of the ISO 639 series of International Standards, not as a universal definition of this concept.
NOTE 2 to entry: Human language also include non-spoken and non-written modalities of human communication which are fully or largely equivalent to human language
[Source: ISO/FDIS 639-4:2009, 3.6 – modified by taking off NOTE 2 to entry which is not applicable in this document and replacing it by the new NOTE 2 to entry]

individual human language
largest set of idiolects that are all interconnected among themselves through (a chain of) high mutual intelligibility and which are socio-politically considered as a unit
NOTE 1 to entry: This definition is intended to serve as a working definition for the purpose of the ISO 639 series of International Standards, not as a universal definition of this concept.
NOTE 2 to entry: Different groups of people may vary with regard to their assessment of a given set of idiolects as forming or not an individual language depending on their assessment of the mutual intelligibility and in particular 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 variety of a language, or a group of closely related languages, is disputed.
[Source: ISO/FDIS 639-4:2009, 3.6 – modified by taking off NOTE 2 to entry which is not applicable in this document and replacing it by the new NOTE 2 to entry]

idiolect
homogeneous means of communications (i.e., a set of linguistic expressions and their meaning, characterized by a coherent system of structural features) capable of coding complex facts and thoughts, used by a given individual human being (the speaker of the idiolect), at a given time, in a given type of situation, and in a given medium
NOTE 1 to entry: Typically, a person has several idiolects of a language at his/her disposal which differ at least in the type of situation and/or the medium that they are applied in.
NOTE 2 to entry: This use of idiolect avoids difficulties of other conceptions where the term idiolect is used in the sense of what is here named a **personal variety**, which is not homogeneous or atomic.

speaker
person making use of systematic linguistic expressions belonging to an idiolect (i.e., expressing him/herself in a language)
NOTE 1 to entry: Speaker serves here as a cover term of all modalities, so it includes the modalities of writing, signing, and other modalities, such as drumming or whistling.

event of language use
instance of language use
event or instance in which a speaker expresses him/herself by means of an idiolect (belonging to a human language)
NOTE 1 to entry: Events of language use can belong to several modalities. In the case of verbal (oral) speaking, this is a speech event, in the case of writing this is an event of producing a written text, etc.

3.2 Dimensions of linguistic variation and language varieties

linguistic variation
NOT: language variant [ISO/FDIS 639-4:2009, 3.14]
range of variation within and between **individual human languages** which, within individual languages, constitutes sub-sets of the language that differ among each other according to external and structural criteria
NOTE 1 to entry: Linguistic variation is seen and can be described as co-existence of different language varieties in one or more of the dimensions of linguistic variation.
[ISO/FDIS 639-4:2009, 3.13 – modified by …]

external criterion for linguistic variation

related properties of sets of idiolects that refers to the speakers by which, or the speech events in which, the idiolects are used
NOTE 1 to entry: Properties belonging to a criterion are all of the same type; i.e. they differ from each other with respect to the same dimension of linguistic variation.

structural criterion for linguistic variation
related properties of sets of idiolects that refer to the structure (i.e. the system) of the idiolects, including in particular phonetic, phonological, morphological, syntactic, lexical, semantic or pragmatic properties

dimension of linguistic variation
one of the types of external criteria according to which subsets of language varieties can be distinguished
NOTE 1 to entry: The dimensions of language varieties are: (1) space dimension (criteria consisting of properties of idiolects referring to geography), (2) time dimension (temporal criteria), (3) social dimension (criteria that refer to social groups to which the speaker belongs by birth, socialization and/or profession), (4) medium (the physical and sensorial channel used), (5) situation (in particular degree of formality), and (6) person (criteria referring to the individual speaker using an idiolect), as well as additional dimensions which systematically usually belong to (6) and that refer to useful (combinations of) criteria, such as: (7) proficiency; (8) particularities of language use (or of performance).

language variety
NOT: linguistic variation
NOT: language variant
largest part (subset) of a language (or of another language variety) that is homogeneous both with regard to some external criterion on a given dimension of linguistic variation and with regard to certain structural criteria.

3.3 Individual language varieties and dimensions of linguistic variation

3.3.1 Space dimension

space
<dimension of language use> set of criteria consisting of properties of idiolects referring to geographical regions
NOTE 1 to entry: This is the dimension of linguistic variation that distinguishes geographically confined dialects or subdialects or supra-regional standard varieties.
NOTE 2 to entry: The space dimension can also cover the aspect of a speaker's dialect which has been influenced by:

- Another dialect, such as High German spoken with Bavarian accent,
- Different stratum of a given dialect or other language, e.g. an Urdu dialect spoken in London influenced by other dialects of the Urdu communities in London and by the English dialect of a given district of London.

dialect
language variety specific to speakers from a certain geographical region
[ISO/FDIS 639-4:2009, 3.8]

standard variety
language variety recognized as acceptable or exemplary by most or all speakers across the geographic area where the language is spoken
NOTE 1 to entry: A standard variety of a language may typically be used in official or public communication and in communication between users of different language varieties. It has often a high degree of status and normalization.
[ISO/FDIS 639-4:2009, 3.15]

3.3.2 Time dimension

time
<dimension of language use> set of criteria consisting of properties of idiolects referring to time

NOTE 1 to entry: This is the dimension that distinguishes historical epochs, periods or smaller spans of time.

period
<dimension of language use> set of criteria specific to a certain span of time which shows a higher degree of internal structural homogeneity of the idiolects belonging to it in comparison to other similarly long spans of time
NOTE 1 to entry: The establishment of periods varies between different experts or experts' communities and depends on their interest or purpose. Begin and end of a period are usually not exact points, so that a period may be characterized by vague delineations or prototypes (e.g.: "the period around the 1880ies" or "the 16th and early 17th century"). The closer to the present moment, the shorter are the periods of a language that can be distinguished due to our more detailed knowledge of the structural features of a language. Periods can span some decades up to a few centuries.

epoch
<dimension of language use> set of criteria specific to a period which shows a higher degree of internal structural homogeneity of the idiolects belonging to it in comparison to other similarly long spans of time
NOTE 1 to entry: Usually epochs of a language are comprised of several periods of the language. They typically span between several centuries up to a few millennia.
NOTE 2 to entry: As with periods, the establishment of epochs varies between different experts or experts' communities and depends on their interest or purpose. Begin and end of an epoch are usually not exact points, so that an epoch may be characterized by vague delineations or prototypes.
NOTE 3 to entry: For several well-studied languages, in particular languages with a long tradition of writing, scholars have established the epochs "old X", "middle X" and "new X", where X stands for the name of the language.

3.3.3 Social group dimension
social group
<dimension of language use> set of criteria consisting of properties of idiolects referring to social groups other than geographically defined groups
NOTE 1 to entry: The social group dimension distinguishes, in particular, sociolects and technolects which may cover several distinct types each.

sociolect
language variety specific to speakers belonging to a certain social group within the society where the language is spoken, either by birth, or by socialization or by acquisition of specialized knowledge

technolect
sociolect that is specific to speakers that have acquired special knowledge of a certain domain or subject

domain
NOT: subject field
field of special knowledge
Note 1 to entry: The borderlines of a domain are defined from a purpose-related point of view.
Note 2 to entry: The delimitation of a domain in terminological entries in standards is usually based on the International Classification for Standards (ICS). In ISO, if the ICS is not suitable in a given case, a domain or subject (3.xxx) should be selected to reflect a purpose, an application or specific requirements.
Note 3 to entry: If a domain is subdivided, the result is again a domain albeit at a higher level of detail.
Note 4 to entry: In IEC (which develops standards in the electrotechnology domain), the usage information related to a term's "specific use" can be a complement to the term but is not necessarily a domain or subject as described in this part of ISO 10241. For further information, see the IEC Supplement to the ISO/IEC Directives, Annex I, *Implementation of the ISO/IEC Directives for the work on the International Electrotechnical Vocabulary (IEV)*[1].
[SOURCE: ISO 1087-1:2000, 3.1.2, modified following ISO 10241-1:2011, 3.3.1]

subject

general topic which is treated or handled in discussion, study, writing, painting, etc.
Note 1 to entry: A subject may touch upon two or more domains.
Note 2 to entry: If a subject is subdivided, the result is again a subject albeit at a higher level of detail.
[SOURCE: ISO 10241-1:2011, 3.3.2, modified – by omitting the reference to Webster]

3.3.4 Modality dimension

modality
<dimension of language use> set of criteria consisting of properties of idiolects referring to the medium
NOTE 1 to entry: When a single linguistic expression is performed several times in different modalities (i.e., it is transferred, re-created, transposed into another modality; e.g., a written text is read out loud, or an oral utterance is transcribed), one can distinguish between a primary and a secondary modality, which are two independent values in the modality dimension.
NOTE 2 to entry: The major modalities are: oral, written, signed, drumming, whistling or using other sound producing methods of the body without speaking, or using instruments such as drums, flutes, etc.

medium
language modality)
<dimension of language use> language variety specific for a certain medium that is used by the speaker when using the idiolects in this language variety
NOTE 1 to entry: An event of using a language can be constituted by more than one modality. Medial varieties include ***spoken language*** (which usually is also multimodal, as gestures, mimic and other similar phenomena almost always accompany speech), ***written language***, ***signed language***, and ***drummed language***, ***whistled language,*** etc.

primary modality
<dimension of language use> language modality used when a given linguistic expression was originally created
NOTE 1 to entry: For example, when a written text is read, the original modality is written language. When a spoken utterance is later transcribed, the original modality is spoken language.

secondary modality
<dimension of language use> language modality in which a given linguistic expression has been transformed, re-created or performed.
NOTE 1 to entry: For example, when a written text is read, the secondary modality is spoken language. When a spoken utterance is later transcribed, the secondary modality is written language.

3.3.5 Situation dimension

situation
<dimension of language use> set of criteria consisting of properties of idiolects referring to the type of situation of language use
NOTE 1 to entry: Situation is the dimension that distinguishes different registers and other types of situations of language use.

register
<dimension of language use> set of criteria consisting of properties of idiolects referring to the degree of formality of language use depending on the situation
NOTE 1 to entry: The major registers are: informal register (i.e. sub-neutral, intimate or casual, some of which according to some conceptions for certain languages constitute, alongside vulgar and familiar etc., sub-registers within the more general informal register), neutral register and formal register (for some languages sub-categories within the formal register may have to be distinguished).

3.3.6 Person dimension

personal dimension
<dimension of language use> totality (i.e. largest set) of idiolects of a language that a given person has at disposal during the lifetime

NOTE 1 to entry: This kind of language variety is often named “idiolect”, but must not be confused with idiolect as defined in this standard. Typically, a person has ***several*** idiolects of a language at his/her disposal which differ at least in the type of situation and/or the medium that they are applied in.
NOTE 2 to entry: Idiolects also evolve over time (e.g., when a new vocabulary is acquired), and for the purposes of this standard it is an open question whether this also gives rise to different idiolects of the same speaker that vary only in the temporal dimension.

person
individual speaker
<dimension of language use> set of criteria consisting of properties of idiolects referring to the individual speaker
~~NOTE 1 to entry: This is the dimension that distinguishes different personal varieties – there is exactly one for each speaker, as each speaker uses the language differently which is to some degree reflected by structural properties of their idiolects.~~

learner
<dimension of language use> set of criteria consisting of properties of idiolects referring to the individual speaker as a learner of a language in one of several different stages of his/her language acquisition process
NOTE 1 to entry: The learner dimension may vary in the degree of nativeness and evolvedness. Thus, the differences in the incipient and intermediate means of communication used by learners even in similar stages of language acquisition/learning may differ greatly from speaker to speaker, also depending on the native language(s) of the speaker and aspects of the knowledge of the speaker. The most salient difference is that of first-language-acquisition and second-language acquisition/learning.

3.3.7 Proficiency dimension
proficiency
<dimension of language use> set of criteria consisting of properties of idiolects referring to the proficiency of the speaker in using the language
NOTE 1 to entry: Proficiency also distinguishes different learner’s varieties. For second language learners for instance, according to some conceptions, one can distinguish at least between ***incipient***, ***intermediate*** and ***advanced*** learner’s varieties.
NOTE 2 to entry: The learner’s varieties in the process of infant first-language acquisition vary according to the theoretical framework.
NOTE 3 to entry: A different kind of ***non-native*** variety are cases when a speaker or author imitates another non-native variety of his or her own native language; e.g. an author writing in a (pseudo) historical period of the language, or speakers imitating a dialect of a region where neither they themselves nor their parents have been socialized, and analogously for sociolects.

3.3.8 Performance peculiarity
performance peculiarity
<dimension of language use> language variety that is specific to situations when speakers demonstrate a certain characteristic particularity in using a language.
NOTE 1 to entry: A performance peculiarity may be congenital or temporarily or permanently acquired. Some can appear only in a certain type of situation (e.g., under stress or strong emotions), others may be constantly present. A speaker may have more than one performance peculiarity at a time. Furthermore, there are degrees of such performance peculiarities up to communication disorders. Well-known peculiarities include stuttering, lisping, dyslexia, etc.

performance
<dimension of language use> set of criteria consisting of properties of idiolects referring to performance peculiarities demonstrated by speakers when using the language

3.4 Documentation of language resources

Language resource
LR
digital resource that provides information about one or more languages
NOTE 1 to entry: Digital resource may also cover resource which may have to be digitized.
NOTE 2 to entry: The language information in the resource may be that of one or more speakers and can be in one or more language varieties as identified and described in this standard.
[SOURCE: ISO 24619:2011 – modified by adding the notes which are necessary for identifying and describing language resources with a higher degree of granularity]

writing system
system for writing a **language** (3.xxx), including the **script** (3.xxx) and character set used
[ISO/FDIS 639-4:2009. 3.16 – modified by adapting the ordering number according to Chapter 3 of this standard]

script
set of graphic characters used for the written form of one or more **languages** (3.xxx)
[ISO 15924:2004, 3.7 and ISO/IEC 10646-1:2003, 4.14 and ISO/FDIS 639-4:2009. 3.17]
NOTE 1 to entry: A script, as opposed to an arbitrary subset of characters, is defined in distinction to other scripts; in general, readers of one script may be unable to read the glyphs of another script easily, even where there is a historic relation between them (see 3.xxx).
NOTE 2 to entry: In certain cases, ISO 15924 provides codes which are not subsumed under this definition. Examples: the codes for aliases and the variant codes.

script code
script identifier
combination of characters used to represent the name of a script
[ISO 15924:2004, 3.8]

conversion
system for representing text in a different **script** (3.xxx) than that in which the text was originally represented
NOTE 1 to entry: The resulting text is also referred to as a “transcription”.
NOTE 2 to entry: The two basic methods of conversion of a system of writing are transliteration and transcription. The use of the terms source script and target script in transliteration is analogous to the terms source language and target language in translation.
[ISO/FDIS 639-4:2009, 3.19 – modified by adapting the ordering number according to Chapter 3 of this standard, and adding NOTE 1 to entry from ISO 15919:2001, 4.1 as NOTE 2 to entry to this standard]

speech-to-text conversion
STT conversion
conversion of speech input to text
NOTE 1 to entry: Speech input can mean the speaking of a speaking taking place or the speech registered in a language resource.
[SOURCE: ISO/IEC 2382-29:1999, 29.0202 – modified by adding the new NOTE 1 to entry and deleting the NOTE1 to entry and NOTE 2 to entry of the entry in ISO/IEC 2382-29:1999]

transliteration
representation of the graphic characters of a source script by the graphic characters of a target script
NOTE 1 to entry: In transcription, pronunciation conventions are of primary importance, while in transliteration, writing conventions are of primary importance.
[SOURCE: ISO 15919:2001, 4.7]

transcription
representation of the sounds of a source language by graphic characters associated with a target language
[SOURCE: ISO 15919:2001, 4.6]

language identifier
language symbol
string of characters assigned to an individual human language for the purpose of uniquely representing it
NOTE 1 to entry: In the language codes of Parts 1, 2, 3, and 5 of ISO 639, each language identifier is composed of two or three letters.
[SOURCE: ISO 639-4:2009, 3.5 – modified by replacing “a linguistic entity” by “individual human language” in line with this standard and deleting NOTE 2 to entry, because it would not be meaningful in this standard]

4 Linguistic variation and language varieties

4.1 Linguistic variation

Not only are individual human languages different one form among another; there is also linguistic variation within each human language. No instance of language use is independent from the language-internal variation. This is also applicable to the description of human language resources that represent instances of use, or other properties of human language.

The identification of different languages is the subject of multipart standard ISO 639 “Codes for the representation of names of languages”. ISO 639 identifies existing (living and extinct) languages, as well as language families and language groups. This standard, in turn, is concerned with the language internal linguistic variation.

Language-internal variation within human languages give rise to language varieties according to distinct dimensions. Although mutual influences exist, each of the dimensions of language varieties is in principle independent from the others. Some varieties can be further differentiated into sub-varieties of the same or other dimensions. Each linguistic manifestation in a given human language (such as a text, an utterance, an entry in a lexical database, etc.) can, therefore, be characterized according to its position in each of these dimensions of language varieties.

The boarders of human languages are sometimes hard to establish. In particular, different groups of people may disagree with regard to their assessment of a given set of idiolects as forming or not an individual language depending on their assessment of the mutual intelligibility and in particular 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 variety of a language, or a group of closely related languages, is disputed. This problem is addressed by ISO 639, in particular ISO 639-3.

Similarly, the borders of language varieties are sometimes hard to establish, and in many cases different varieties being distinguished on the same dimension (e.g., different dialects) do overlap (that is, there may be idiolects that fulfil all criteria to belong to both varieties).

In addition, human languages and their language varieties are under constant gradual – sometimes fast – change so that:

- a given language variety may become considered as a language and vice versa (to be covered in ISO 639);
- linguistic expressions and features belonging to a given language variety at one point may change their position with respect to one or more of the dimensions of language varieties (e.g., markedly informal expressions may become acceptable even in formal contexts, or regional expressions may spread over large parts of the geographical territory where the language is spoken).

This standard focuses on a framework for the identification and description of language-internal variation, that is, of language varieties and sub-varieties. It does not aim at establishing all language varieties of all languages – even at a given point in time this would be an immense list. But it exhausts the ***types*** of descriptors that may be needed for characterizing the status of a given instance of language use (e.g., an utterance) or language resource. It does so by describing all the dimensions in which languages can vary internally, and indicates the major resulting varieties that typically occur in natural human languages.

Technically, languages are conceived in this standard as sets of idiolects (means of communications), and language varieties as subsets. The language is classified simultaneously into different varieties according to different kind of certain criteria. The resulting varieties in each classification can overlap, and they can be sub-classified into smaller varieties, again according to certain criteria.

Each language variety is characterized by structural properties (of the sound system, the morpho-syntax, the lexicon or semantic system) and at the same time certain external properties (to be spoken in a certain geographical area, at a certain time, etc.). The organization of these criteria into a few major types constitutes the different dimensions of linguistic variation.

4.2 Dimensions of linguistic variation

Linguistic manifestations such as language resources and the events of language use themselves can be characterized according to the following dimensions of linguistic variation:

1) Space (dialects, sub-dialects as well as supra-regional standard varieties)
2) Time (epochs, periods, stages)
3) Social group (sociolects of several different types and technolects)
4) Modality (spoken, written, signed, whistled, drummed, etc.)
5) Situation (registers of different formality, also for motherese and similar varieties)
6) Individual speaker (often mot precisely called “idiolects”)
7) Proficiency (for learners’ varieties of different stages)
8) Performance (peculiarities up to certain communication disorders)

A complete characterization of the position of any given event of language use and any language resource with regard to linguistic variation would state the position of the event or resource with respect to each of these eight dimensions.

Language varieties of some dimensions may be complex within themselves by having more sub-varieties of the same or possibly even other dimensions, which implies that there may be more than one value on these dimensions for a given event of language use or a given linguistic resource. If there are two values on the same dimension, then usually one value is more specific and the other broader (e.g., a recording of a speaker using a Norfolk dialect can also be characterized as belonging to the broader East Anglian dialectal variety of English).

There may also be mixed events or language resources that contain several instances of language use that belong to different varieties according to the same dimension (for instance a dialogue between speakers that use different dialects, or a dictionary that covers several dialects, sociolects etc. of a language). In such cases, all (groups of) varieties at hand will have to be identified, and if possible, the respective parts of such a resource will have to be related to their respective varieties (e.g., the different participants and accordingly the different time segments of such a dialogue; or the different entries in such a dictionary).

Another special case are language resources where the language use is “non-native” in the sense of deliberately imitating another than the native variety of the speaker. For instance, a speaker of a dialect X tries to imitate speech of another dialect Y, or an author makes use, e.g. as a stylistic device, of language as conceived as typical for a certain historical period of time, or a certain social group to which the writer does not belong. The latter is covered by the proficiency dimension in this standard.

4.3.1 The space dimension of linguistic variation

According to the space dimension, a language can be differentiated into different 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: East Anglian (English dialect)
EXAMPLE 2: United Kingdom Standard English (Supra-regional standard variety)

When determining the dialect of a speaker in a given situation, what is mainly considered is the geographical region of the socialization of the speaker, that is, where the speaker grew up and where the speaker’s parents grew up. If these factors are heterogeneous (due to, e.g., migration, parents from different regions), one may be able to identify the major dialect to which the language use of the speaker most strongly resembles, alongside with minor dialectal influences (e.g., 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).

EXAMPLE: “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. How many and which dialects are to be distinguished on a given level of specificity is often debated between specialists, and so are the technical names and the borders of the dialects.

In 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 accepted as “standard” by most or all speakers across the whole geographical area of the language. 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 (e.g., as an ‘accent’), even with speakers that do not have strong competence in the original local dialect.

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

EXAMPLE: A certain Urdu dialect spoken in London (with influence from the English north of the Thames)

When speakers speak several dialects and/or the standard variety more or less fluently, it may have to be determined for each event of language use which of the dialects or standard variety has been used. For some purposes it may additionally be useful to indicate which other languages and which other dialects of the language in question the speakers are able to speak.

4.3.2 Time dimension of linguistic variation

According to the time dimension of linguistic variation, a language can be differentiated into different historical epochs, periods or stages. They may be named after the eras of political organization, of rulers, or cultural, social or economic development. Epochs may comprise distinct periods, which in itself may comprise stages.

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

EXAMPLE 1: Early Middle English
EXAMPLE 2: Victorian English

Neither the epochs, let alone the periods, are identical in their temporal expansion between languages, and sometimes not even between different dialects within one language.

The establishment of periods varies a lot between different experts and depends on their interest or purpose. Begin and end of a period are usually not exact points, so that a period may be characterized by vague delineations or prototypes (e.g.: “the period around the 1880ies” or “the 16th and early 17th century”). The closer to the present moment, the shorter are the periods of a language that can be distinguished due to our more detailed knowledge of the structural features of a language. Still, periods typically span some decades up to a few centuries.

4.3.3 Social group dimension of linguistic variation

According to the social group dimension of linguistic variation, a 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, or gender.

The number and specificity of sociolects that need to be distinguished varies very much from language to language and reflect the social structure und in particular the social segregation of the society in which the language is used. In some small egalitarian societies there may be no social group variation at all (i.e., just one general neutral sociolect), other small societies may only strongly differentiate between two (i.e. a male and female) genderlects. On the other end of the spectrum we have complex societies with different language varieties for each of a number of social strata, and nowadays an even larger number of technolects or jargons for different professional groups.

EXAMPLE 1: Male Aweti (genderlect)
EXAMPLE 2: Academic English (sociolect)
EXAMPLE 3: Discussion of medical doctors among themselves (technolect)
EXAMPLE 4: Automotive workers in a car factory (jargon)

In delimitating, describing and naming social groups, there is often a disagreement between experts. Not all social differentiation necessarily leads to distinctive sociolects.

Technolects and jargons refer to the specialized purpose language varieties characteristic for the communication in certain domains or subjects. In naming and describing the groups using technolects, there is a constant evolution in the sciences, technologies and economies of the world. Therefore, the categories under the social group dimension are an open list.

4.3.4 Modality dimension of linguistic variation

According to the modality dimension of linguistic variation, a language can be differentiated into different modalities which refer to the sensorial channel what is used in an event of language use, and consequently in the language resources that represent these events of language use. The events and resources can belong to the following modalities:

- Multimodal language use (in particular spoken language combined with gestures; this is the most common type of language use)
- Spoken language use events (a pure audio recording, or language use on the telephone)
- Written language use
- Signed language use (for sign languages; this is different from the use of –even conventionalized– gestures)
- Use of augmentative and alternative communication (AAC) such as symbol systems (Bliss symbols and the like, as far as these have the communicative power comparable to that of natural human languages)
- Language use in the haptic modality (in particular when communicating with persons who are both blind and deaf)

- Language use of other acoustic modalities performed with only the speaker's body (such as hummed and whistled languages)
- Language use of other acoustic modalities performed with the help of external tools (in particular music instruments such as using drums or flutes)

The speech modality is usually in fact the multimodal modality, because speech is often accompanied by certain kinds of non-verbal communication, such as mimics and gestures, which may necessitate specification.

Besides from that, an event of language use recorded in a language resource may comprise more than one modality, e.g. sign language use between deaf speakers (signers) intermitted with spoken language.

There may be a dominant modality and one or more additional modalities, such as a written text accompanied by bliss symbols.

When a single linguistic expression is performed several times in different modalities (i.e., it is transferred, re-created, or transposed into another modality; e.g., a written text is read out loud, or an oral utterance is transcribed), one can distinguish a primary and secondary modality, which are two independent values in the modality dimension.

EXAMPLE 1: When a written text is read, the original modality is written language, and the secondary modality spoken language.
EXAMPLE 2: When a spoken utterance is later transcribed, the original modality is spoken language, the secondary modality is written language.

Within the written modality, one can further specify which writing system, which script, which orthography etc. are being used. The identification of this aspect is already covered by other standards.

4.3.5 Situation dimension of linguistic variation

According to the situation dimension of linguistic variation, a language can be differentiated into different registers.
The most relevant types of situations of language use distinguish themselves by the *degree of formality* or of respect, from informal to formal, and includes the neutral register (expressions that are appropriate in all kind of situations). Hence, the major primary registers are:

1) ***informal*** (= *sub-neutral*, *intimate*, *casual*, according to some conceptions for some languages contain sub-register within the more general informal register, e.g. *vulgar* and *familiar* etc.),
2) ***neutral*** (appropriate in informal and formal situations alike), and
3) ***formal*** (again, for some languages sub-categories within the formal register may have to be distinguished, for instance when addressing a king of religious leader).

Other registers that can be distinguished for the purposes of this standard are

4) ***motherese*** (a register used in particular by parents to speak to young children), and

5) ***foreigner-talk*** (a register used to speak to not fully fluent adult language learners), and the like.

In some historical cases the latter register has developed into a widely used form of communication which became a pidgin, which is on the boarder of being a widely used register and/or sociolect or a language of its own. From pidgins can then in turn develop creole languages if they are acquired by children as their major language.

In some (in particular east and south-east Asian) languages exists a more sophisticated system of registers that is closely intertwined with the social dimension, because the style, vocabulary and even grammar changes according to the relative and absolute social position (including age) of both, the speaker and the addressee.

One further type of variation belongs systematically here because it depends on the kind of situation of language use: the genres. Again, which and how many genres need to be distinguished can vary very much between languages. Many languages differentiate between at least these:

6) ***epic/prosa*** (regular language formally not constraint other than by factors covered above in this or other dimensions),
7) **lyrics/poetry** (language organized according to primarily aesthetical criteria, also often used when singing), and
8) **ritualistic language** (in ritual, spiritual, magical or religious contexts).

It is an empirical question whether more specific categories are needed in a given language. In principle this standard is only concerned with linguistic variation that affects the language structure (which includes the lexicon), and not with differences what concern merely different styles of language use, and the difference between, say, a crime thriller novel and a history book is arguably not affecting the underlying linguistic system (even in the lexicon, different frequencies of words are a different phenomenon from words which are ungrammatical in a certain situation). Still, the characterization of a more specific genre may be useful for certain purposes (in particular research).

4.3.6 Individual person dimension of linguistic variation

According to the personal dimension of linguistic variation, a language can be differentiated into different ***personal varieties***.

There is exactly one personal variety for each speaker, because every person speaks his/her own personal variety, which sometimes called ‘idiolect’. The idiosyncrasies that characterize individual person’s language varieties can concern all levels of the linguistic structure, in particular the lexicon.

Within the personal variety of a speaker there are usually many different idiolects in the sense of this standard (which are homogeneous, that is, characterized by a coherent system of structural features), which differ at least in the type of situation and/or the medium that they are applied in. That is, some of a speaker’s idiolects belong to

different registers (formal vs. informal) or to different modalities (e.g., written vs. spoken/multimodal), and sometimes even to different dialects of sociolects.

4.3.7 Proficiency dimension of linguistic variation

According to the proficiency dimension of linguistic variation, a language can be differentiated into different stages of learning (learner's varieties). These are specific to learners of a language in one of several different stages of their language acquisition process. They vary in their degree of nativeness and evolvedness.

Systematically, this dimension could be covered by the person and time dimensions – it is a matter of fact that a given language learner (child or adult) at a given point in time is acquiring the language and has reached a certain stage. Still, for the purposes of, for instance, adapting language technology or resources for teaching the language, it may be necessary or useful to state the stage of learning as a different dimension.

The differences in the incipient and intermediate means of communication used by learners even in similar stages of language acquisition/learning may vary greatly from speaker to speaker, also depending on the native language(s) of the speaker and other aspects of the knowledge of the speaker. The most salient difference is that of first-language-acquisition and second-language acquisition/learning.

The learner's varieties in the process of infant first-language acquisition vary according to the theoretical framework. Sometimes used concepts are, e.g., "holophrastic stage", "two-word-stage", etc.

For second language learners, according to many conceptions, one can distinguish at least between the following learners' language varieties:

1) ***incipient***,
2) ***intermediate***, and
3) ***advanced***

A different kind of ***non-native*** variety can be found in cases when a speaker or author imitates another non-native variety of his or her own native language; e.g. an author writing in a (pseudo) historical period of the language, or speakers imitating a dialect of a region where neither they themselves nor their parents have been socialized, and analogously for sociolects.

4.3.8 Performance peculiarity dimension of linguistic variation

Following the point of view of health impairments, a communication disability is a functional problem for a person. In the context of this standard, communication disabilities manifest in the form of performance peculiarities.

Communication disabilities (i.e. communication anomalies or communication disorders) can be

- congenital or
- temporarily acquired or

- permanently acquired.

The degrees of a communication disability can be comparatively light, medium or strong. A given communication disability may trigger or influence other communication disabilities, e.g. light stuttering may make a person hush or negatively impact the person's fluency of speech.

Communication disabilities are part of a multidimensional system and can have different causes, such as:

- health impairments, which comprise a loss or abnormality of physiological, psychological, or anatomical structure or function; in connection with communication disabilities they comprise in particular
  - anomalies or disorders of the nervous system (including the brain), respiratory-phonatory system, oropharyngeal system, etc.
  - impairment of the senses,
- psychological conditions/factors impacting behavior or mind or cognition,
- externally induced impairments that may result in communication disabilities include:
  - physical impact of accidents, surgery, natural forces, etc.
  - environmental effects, including natural environment phenomena taken up by the senses, social environment influencing mood, cognition, behavior, etc.
  - intake or application of medicines, drugs, chemicals, etc.
  - excessive radiation,
- functional impairments/disabilities which cannot be subsumed under the above.

The above-mentioned causes may or may not result in communication disabilities. However, they do result in performance peculiarities and can cross-influence each other.

With respect to communication disabilities the above-mentioned causes materialize in the form of performance peculiarities of

- speech
- language
- voice
- behaviour.

The degree of the performance peculiarity can range from a light communication anomaly to a communication disorder.

Speech-related performance peculiarities refer to components of speech production, which include:

- phonation
- resonance
- fluency
- intonation
- variance of pitch
- aeromechanical components of respiration.

Language-related performance peculiarities refer to the composition of language structures, which include:

- phonology
- supra-segmental features (e.g. changing phones according to the rules of a language)
- morphology
- syntax
- grammar rules
- semantics
- pragmatics.

Voice-related performance peculiarities refer to conditions involving abnormal pitch, loudness or quality of the sound produced by the larynx and thereby affecting speech production.

Behaviour-related performance peculiarities may be due to any or a combination of the above-mentioned causes. They can impact

- speech production, language structure composition and voice generation,
- the ability to grasp communication content, thus resulting in ostensible cognitive disorder affecting the communication.

They can also materialize in anomalies or disorders of non-verbal-communication or non-attentiveness hindering communication. Sometimes they are due to strong individual habits, such as a snoring sound or tongue-clicking during the speaker's speech, not necessarily due to any of the above-mentioned causes.

5 Indication of language varieties

5.1 Scope of the standard

This standard focuses only on the identification and description of language varieties, not on the genera, formal or technical aspects of the description of human language resources, which are covered by other metadata frameworks.

For instance, for the general description of a language resource, it is recommended to apply at least the metadata of the OLAC Metadata standard http://www.language-archives.org/NOTE/usage.html, which provides an application of the Dublin Core metadata element set as defined by the Dublin Core Metadata Initiative (DCMI). http://dublincore.org/documents/dcmi-terms/

The Component Metadata Infrastructure (CMDI) provides a best practice guide for the sake of technical and content interoperability between language resources as well as of their sustainability. http://www.clarin.eu/content/cmdi-best-practice-guide

The Metadata Infrastructure for Language Resources and Technology lists the metadata needed for the formal description of language resources as follows: (s. CLARIN Registry Requirements http://www-sk.let.uu.nl/u/D2R-4.pdf)

- External metadata for language resources
- Lexicon Metadata: macrostructure and microstructure

- Metadata set for text corpora
- Metadata set for speech resources
- Metadata set for multimodal resources
- Metadata set for tools

All major metadata sets for the formal description of human communication resources comprise the identification of the language of the resource following the multipart standard ISO 639 Codes for the representation of names of languages.

This standard specifies which type of descriptors are needed to exhaustively account for the place of an event of language use or a language resource in the multidimensional space of variation within a language. Therefore, this standard represents an extension to the major metadata standards such as the OLAC or CMDI metadata formats.

In particular, it could be employed as the basis for a coherent system of extensions to the widely used Best Current Practice 47 that specifies “Tags for Identifying Languages”. That recommendation already foresees the identification of language varieties (in particular, dialects, scripts and regions), but this is far from complete. It has also room for defining extensions to sub-tag elements. These may be used to implement this standard, e.g. by establishing one extension for each dimension.

Currently, no concrete values to be used within the framework here defined are proposed, but clearly, a mechanism to register these values is needed and should complete this standard in later updated versions. Authorization authorities will be needed for registering concrete values and their representation (code-elements or tag-elements), in particular for those dimensions that are for all practical purposes open lists, such as dialects (for all languages), periods (for all languages), and sociolects. According to this standard, the modalities and registers as well as proficiency learner’s varieties and communication peculiarities are smaller and limited lists. The individual dimension of linguistic variation is covered by identifying the speaker(s).

5.2 Indication of individual language varieties

Any given event of language use (represented in a language resource) belongs simultaneously to a certain dialect, to a certain period, to a certain modality (is e.g. written, or oral), to a certain sociolect, etc. Therefore, the values for each of these dimensions can be stated side by side.

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

For the finding and re-use of language resources, the specification/identification of linguistic varieties may be crucial, even if they were not in the focus of the creators of the resources at hand. Therefore, the position of a given language resource ideally should be made explicit for each of the dimensions of language variety.

However, not in all cases it may be possible, or even necessary, to indicate the respective varieties according to all dimensions. In the course of the description and identification of the language variety at hand, omission of a dimension should also be made explicit, for instance by marking them as “unspecified”.

EXAMPLE:
[dialect:] unspecified

If the specification/identification of a dimension of language variety lacks confirmation, e.g. by being assumed or inferred, the (in)certainty of the specification/identification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred” or similar. This is then a property that refers to the respective statement of the language variety, not to the event of language use or the language resource in question.

The fact that a language resource contains several instances of language use that belong to different varieties according of the same dimension (so as a dialogue between speakers that use different dialects) should be clearly stated.

In case that the language used in a language resource is deliberately chosen to imitate that of another variety, 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.

5.2.1 Indication of the space dimension of linguistic variation

The space dimension of an event of language use, or a language resource that represents an event of language use, is specified by identifying the dialect within which or location where the speaker grew up, and ideally where the dialect / 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 linguistic variety of a region, that name should be used. For the sake of international communication, the English name of the dialect should be used (possibly in addition to the original name of the dialect in the language itself, or in another meta-language).

If the specification/identification of the dialect lacks confirmation, the (in)certainty of the specification/identification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”, or similar.

EXAMPLE 1:

[Dialect:] Irish English; [Subdialect:] unidentified
EXAMPLE 2:
[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 for the speaker of the event of language use in a given language resource.

EXAMPLE:
[language:] Urdu
[Dialect:] London; [Influence by language:] English

5.2.2 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 period or epoch of an event of language use, that name should be used. It is possible to make several indications, starting with the larger epoch and becoming more specific. In principle, the most specific value implies the others, but these may still be useful for other purposes. Also in periods of transition it may be useful to state if a given event or resource can be more clearly grouped into one period than another.

For the sake of international communication, the English name of the given period or epoch should be used (in addition to the original name of the period or epoch in the native language, or another meta-language).

EXAMPLE 1:
[Epoch:] Modern English; [Period:] Victorian English

If the specification/identification of the period or epoch lacks confirmation, the (in)certainty of the specification/identification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”.

EXAMPLE 1:
[Dialect:] Irish English; [Subdialect:] unidentified
EXAMPLE 2:
[Dialect:] Irish English; [Subdialect:] Dublin English; [Status of subdialect:] assumed

NOTE: The date and the time of the creation of a language resource (by e.g. digitizing the recording of a communication event) is part of the general metadata of the language resource and thus may differ from the date of the event of language use itself. The latter is relevant for the identification of the relevant temporal variety. In some cases, several dates have to be distinguished (e.g., the digitization of a 19th century re-edition of a 17th-century text; in such a case the period to be indicated according to this standard would be the 17th century, which is when the original event of language use, the writing of the text, happened).

5.2.3 Indication of the social group dimension of linguistic variation

The sociolect(s) used in an event of language use should be specified as exactly as possible. In general, this is done by stating the social stratum or sub-community in which the speaker was socialized – ideally together with the speaker's parents' social groups. If there is an established name for the sociolect of an event of language use, that name should be used. Depending on the topic of the event of language use or the linguistic resource, it may be most relevant to state the education and/or occupational group of the speaker(s).

For the sake of international communication, the English name of the sociolect should be used (in addition to the original name of the sociolect in the language itself or in another meta-language).

EXAMPLE 1:
[Sociolect:] African American Vernacular English
EXAMPLE 2:
[Sociolect:] Middle class English; [Sociolect/Education:] Academic
EXAMPLE 3:
[Sociolect/Technolect:] (English) business speak

If the specification of the sociolect lacks confirmation, the (in)certainty of the specification/identification should be made explicit by stating "unconfirmed", or more specifically "assumed" or "inferred".

EXAMPLE 1:
[Sociolect:] Middle class English; [Status of sociolect:] inferred

5.2.4 Indication of the modality dimension of linguistic variation

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

For the sake of international communication, the English name of the modality should be used (in addition to the original name of the modality in the language itself or in another meta-language).

EXAMPLE 1:
[Modality:] Spoken/Multimodal

If the specification/identification of the modality lacks confirmation, the (in)certainty of the specification/identification should be made explicit by stating "unconfirmed", or more specifically "assumed" or "inferred".

EXAMPLE 1:
[Modality:] Multimodal; [Status of Modality:] inferred
(e.g., in the case of an audio recording where it can be inferred that the speakers also saw one another)

In the case of more than one modality represented in a given Language resource, they should all be stated together with the indication of primary and secondary modality, if possible or useful.

EXAMPLE 1:
[Modality/primary:] Written; [Modality/secondary:] Spoken
(e.g., in the case of a written text read aloud)
EXAMPLE 2:
[Modality/primary:] Spoken; [Modality/secondary:] Written
(e.g., in the case of a transcript of a dialogue)

Sometimes written communication events are transcribed into a different writing system that the original writing system of the language variety. Frequently non-written language varieties are transcribed. For this transcription different writing systems and different transcription conventions may be applied. For the identification of regular writing systems the international standard ISO 15924:2004*, Information and documentation — Codes for the representation of names of scripts* shall apply.

For different transcription conventions the name of the given transcription system should be used. For the sake of international communication, the English name of the transcription system of a given modality should be used (in addition to the original name of the transcription).

If the specification/identification of the modality dimension of language variety – or of its transcription systems – lacks confirmation, the (in)certainty of the specification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”.

5.2.5 Indication of the situation dimension of linguistic variation

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

For the sake of international communication, the English name of the modality should be used (in addition to the original name of the modality in the language itself or in another meta-language).

EXAMPLE 1:
[Register:] Informal, [Register/Context:] Familiar
EXAMPLE 2:
[Register:] Neutral; [Register/Addressee:] Child-directed speech

If the specification/identification of the register lacks confirmation, the (in)certainty of the specification/identification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”.

EXAMPLE 1:
[Register:] Formal; [Status of Modality:] inferred

5.2.6 Indication of the person dimension of linguistic variation

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

This is usually covered by other parts of the formal description of language resources (i.e. metatada categories). In case that this is not the case, it can be done as part of the framework of this standard.

EXAMPLE 1:
[Person:] John Doe

If the specification/identification of the person lacks confirmation, the (in)certainty of the specification/identification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”.

EXAMPLE 1:
[Person:] John Doe; [Status of Person:] assumed

5.2.7 Indication of the proficiency dimension of linguistic variation

The proficiency (type of language acquisition process and stage of learning) should be specified as exactly as possible. If there is an established name for the proficiency of an event of language use that name should be used.

For the sake of international communication, the English name of the modality should be used (in addition to the original name of the modality in the language itself or in another meta-language).

EXAMPLE 1:
[Acquisition:] Second language acquisition; [Proficiency:] Beginner
EXAMPLE 2:
[Acquisition:] First language acquisition; [Proficiency:] Two-word-stage

If the specification/identification of the stage of language learning lacks confirmation, the (in)certainty of the specification/identification should be made explicit by stating “unconfirmed”, or more specifically “assumed” or “inferred”.

EXAMPLE 1:
[Acquisition:] Second language acquisition; [Acquisition/Context:] School; [Proficiency:] Intermediate; [Status of proficiency:] inferred

5.2.8 Indication of the performance peculiarity dimension of linguistic variation

The performance peculiarity dimension of linguistic variation should be specified as exactly as possible. If there is an established name for the performance peculiarity or communication disorder of an event of language use that name should be used.

For the sake of international communication, the English name of the performance peculiarity or communication disorder should be used (in addition to the original name of the performance peculiarity).

EXAMPLE 1:
[Particularity:] stutter
EXAMPLE 2:
[Particularity:]

If the specification/identification of the performance peculiarity or communication disorder of an individual speaker's variety lacks confirmation, the (in)certainty of the specification/identification should be made explicit by stating "unconfirmed", or more specifically "assumed" or "inferred".

EXAMPLE 1:
[Particularity:] lexical disorder; [Status of particularity:] assumed

ANNEX A

Recommendation on software and content development principles 2010

Formulated at the ICCHP 2010[4] and endorsed by ISO/TC 37 and other technical committees

Purpose

This recommendation addresses decision makers in public as well as private frameworks, software developers, the content industry and developers of pertinent standards. Its purpose is to make aware that multilinguality, multimodality, eInclusion and eAccessibility need to be considered from the outset in software and content development, in order to avoid the need for additional or remedial engineering or redesign at the time of adaptation, which tend to be very costly and often prove to be impossible.

Background

In software development, globalization[1], localization[2] and internationalization[3] have a particular meaning and application. In software localization they have been recognized as interdependent and of high importance from a strategic level down to the level of data modelling and content interoperability.

In 2005 the Management Group of the ITU-ISO-IEC-UN/ECE Memorandum of Understanding on eBusiness standardization adopted a statement (MoU/MG N0221), which defines as basic requirements for the development of fundamental methodology standards concerning semantic interoperability the fitness for

- multilinguality (covering also cultural diversity),
- multimodality and multimedia,
- eInclusion and eAccessibility,
- multi-channel presentations,

which have to be considered at the earliest stage of

- the software design process, and
- data modelling (including the definition of metadata),

and hereafter throughout all the iterative development cycles.

The above requirements are a prerequisite for global content integration and aggregation as well as content interoperability. Content interoperability is the capability of content to be combined with or embedded in other (types of) content items and to be extensively re-used as well as re-purposed for other kinds of eApplications. In order to achieve this capability, software must support these requirements from the outset. The same applies to the methods and tools of content management – including web content management.

Recommendation

Software should be developed and data models for content prepared in compliance with the above-mentioned requirements to facilitate the adaptation to different languages and cultures (localization) or new applications (re-purposing), the personalization for different individual preferences or needs, including those of persons with disabilities. These requirements should also be referenced in all pertinent standards.

[1] **Globalization**) refers to all of the business decisions and activities required to make an organization truly international in scope and outlook. G11N is the transformation of business, processes and products to support customers around the world, in whatever language, country, or culture they require.

[2] **Localization** is the process of modifying products or services to account for differences in distinct markets. Therefore, L10N is an integral part of G11N, and without it, other globalization efforts are likely to be ineffective. The interdependence of G11N and L10N has also been coined **glocalization**.

[3] **Internationalization** is the process of enabling a product at a technical level for localization. An internationalized product does not require remedial engineering or redesign at the time of localization. Instead, it has been designed and built from the outset to be easily adapted for a specific application after the engineering phase.

[4]**ICCHP**: International Conference on Computers Helping People with Special Needs