Towards Common Language Codes

User Requirements

Language Codes

Characters designating languages
E.g., ar or ara or az for Arabic

Importance of common code
Use of tools for hyphenation, spell checking, search, knowledge management tools, machine translation

Agenda

Uses
Requirements
Current Standards
Options
Concerns
Questions

Comprehensive List of Uses of Language Codes

Marking materials and/or specific text/voice/audio for appropriate use of tools
Hyphenation
Spell checking
Search (limit)
Machine translation
Knowledge management and visualization tools
Other language processing

References to items (books and resources, code for language of summaries, text, librettos, table of contents, language skills/use in census or database, personnel; needed in statistical analysis)

Comprehensive List of Requirements for Language
Codes

- Common code for application of tools
  - Decreases development and maintenance costs
  - Increases number of tools available
  - Enables increased use and integration of COTS products
- Applicability across required languages
  - Industry-required
  - Larger group
- Additional information for some applications
  - Country or regional variation (“localization”)
    - Formatting for dates, time, etc.
    - Spell checking
  - Dialects
  - Register (how to handle chat language? Register or modality? How to handle jargons and controlled languages? Differences between dialects and registers in Arabic?)
  - Writing system)
    - (e.g., native text, type of transcription system if transcription is used)
  - Modality (voice, text, sign language?)
  - Time?
  - Denotation of exact meaning of code

Current Standards for Language Codes

*ISO 639*
  - 2-letter codes
    - two 3-letter codes
*ANSI/NISO*
*IANA*
*SIL*
*Voice XML*
*Microsoft*
*Apple*
*IBM*
*OpenType*

**Programmer created**

Concerns with Current Language Code Standards

- Multiple standards
- Other standards (e.g., JAVA, LINUX, etc.) citing limited 2-letter standards
- Organizations already heavily invested in tools with certain standards
- Conflicts in ISO 639
  - 2-letter
  - 3-letter (two versions)
- Coverage
  - 6000+ languages
  - Expanding industry use
Rapidly expanding native use
Detail: dialect, registry, modality, transcription, time
Difficulty of adding to ISO standards
Difference in definitions of categories
Different needs for information; different definitions
Exchange of data resulting in problems with private use space
Difficulty in XML of handling multiple values for LANG

Cost of Not Having Common Language Codes
Need for many converters, including user definable converters, language code identification
Possibility that fewer tools will be developed that could take advantage of language tags, due to the cost of dealing with the complexity
Discontinuity of codes across speech and text and workflow
Less interoperability or more complexity across systems using different language codes

Options
Develop new 4-letter code for ISO 639
Use of only 4-letter ISO code
Use of 2-letter codes, default to 3-letter, then default to 4-letter
Use RFC3066 [superceding RFC1766] language tags to use 2-letter tags, then 3-letter tags, then apply to IANA (could be 4 codes)
IANA perhaps to use SIL codes where they exist
Use of ISO 639 with private-use codes defaulting to SIL (x-SIL-code)
Adoption of SIL codes
In total as new ISO 639
For adding languages to ISO 639 where codes do not conflict; may be temporary
Use systematic extensions to show dialect, country, register, modality, transcription, etc.
Provide incremental or complete solutions

Questions
Is it better to provide a near-term solution and a long term solution, or to just provide a long term solution?
What short term solutions should be considered?
- Adding to ISO 639 for the languages needed now?
  - Are 50 documents easily available (on the web, via LOC, etc.)
  - Combining codes (defaulting from one code to another)
  - Which code: SIL? IANA? Linguashere Registry

More Questions
What additional information needs to be provided for on a voluntary basis?
More Questions

▲ How should changes be handled concerning designations as language or dialect (e.g., upgrades)?
▲ How should other changes be handled?
▲ How do we handle names and codes that can refer to more than one thing (e.g., language)?
▲ How do we get consistent definitions of categories?
  ▲ What principles should be followed for mapping, etc.? (PC)
▲ How do we get consistent use of definitions?
▲ How do these decisions impact tools developers?
  ▲ What is the impact of having synonyms?
  ▲ What is the impact of having an inconsistent number of extensions?
  ▲ Would certain decisions impact user goals of having inexpensive tools that work on a variety of materials and/or systems?
▲ Other?

Will this Work for the Short Term?

▲ Compile a list of languages now required that are not in ISO 639
▲ Attempt to obtain 50 documents in each language
▲ For languages where such documents can be easily obtained, submit formal ISO request
▲ Use language code from Linguasphere or SIL, if that code has not already been used in ISO 639. Otherwise provide new language code not in Linguasphere or SIL and not in ISO