Towards a new improved setting for the ISO 639 standards: the FROLIC approach

Workshop on Language Identifying Codes
Newcastle, 2013-02-09

SEBASTIAN DRUDE
The Language Archive – Max Planck Institute for Psycholinguistics

TOPICS

1. Introduction: The FROLIC project proposal
2. Organizational & administrative strategy
3. Technical strategy
4. Content-related FROLIC proposals

TOPICS

1. Introduction: The FROLIC project proposal
2. Organizational & administrative strategy
3. Technical strategy
4. Content-related FROLIC proposals

1) The FROLIC project proposal

- 2011: ISO-TC37/SC2 plans to revise the technical and organizational setting for ISO 639
- Gerhard Budin proposed to involve CLARIN
- 2012: EU & AUS brainstorming meeting in Brussels
- Europa & USA: Opportunity for a new project in NEH & DFG Bilateral Digital Humanities program

  2008—2011: *RELISH* (Rendering Endangered Languages Lexicons Interoperable through Standard Harmonization)
- 2012-09: Proposal: FROLIC (Framework for the Organization of Language Identification Codes)
  U Frankfurt, ILIT (U-Michigan, LinguistList), TLA @ MPI-PL

1) The FROLIC project proposal

Scope and goals of FROLIC:

- With input from external experts and with global collaboration: develop clear criteria and a conceptual framework for ISO 639-3, -5 & -6 (➔ contributing to an improved ISO 639-4)
- Apply the framework to four selected language families: *NE Caucasian*, *Turkic*, *Lower Cross*, *Berber*
- Use the results to tag lexicons of some languages in these families (enriching LEXUS & LEGO res.)
- Collaborate internationally to improve ISO 639

TOPICS

1. Introduction: The FROLIC project proposal
2. Organizational & administrative strategy
3. Technical strategy
4. Content-related FROLIC proposals

2) Organizational strategy

Current situation:

- ISO TC37 / SC 2: chair is Debbie Garside
- ISO TC37 / SC 2 / WG1: convener position vacant
- Joint Advisory Board, SD ∈ ~7 voting members
- Unclear criteria, procedure & expert involvement

Already agreed in ISO TC37 / SC 2:

- New organizational setting will be installed
- Improved criteria & guidelines (ISO 639-4)
- A Scientific Advisory Board of some 7 to 20 international experts is needed

2) Organizational strategy

Scientific Advisory Board:

- Experts on language cataloging or overview work worldwide or for a certain wider region
- They contact experts on groups of languages
- Respected, not controversial persons
- PRAGMATIC, driven by the wish to come to agreements and practical solutions, i.e.:
- No “barkers”, not driven by personal interests
- No perfectionists or idealists defending an ultimate truth, accepting controversial views

2) Organizational strategy

What is needed from us:

- Support new ISO TC 37 / SC 2 / WG 1 chair
- Select & nominate experts for SAB
- Make pragmatic proposals for ISO 639-4
- Support a new registration authority (CLARIN?)
- Accept that not all expectations can be met:
- ISO 639-3 will have to be revised wrt. inexistent languages, double codes referring to one language, missing languages, etc., but impossible to improve the choice of individual codes already assigned

TOPICS

1. Introduction: The FROLIC project proposal
2. Organizational & administrative strategy
3. Technical strategy
4. Content-related FROLIC proposals

3) Technical strategy

Current situation:

- Separate databases with varying quality
- ISO TC 37 / SC 2 already decided changes

ISO plans (G. Budin):

- One integrated database for all parts
- Designed based on the ISOcat & RELcat model
- Entries can be proposed by a large community
- Standardized entries are approved by SAB
- DB originally planned to be installed in Vienna in cooperation with TLA (host of ISOcat & RELcat)

3) Technical strategy

Two aspects of database setting:

1. Design, technical development & implementation
   - short-term project money is an option
   - data categories have to match requirements
2. Maintenance of DB & underlying technology
   - some sustainable financing model is needed

- Worrying: in ISOcat, the actual standardization of categories is very slow due to over-commitment of members of the committees – who pushes?

3) Technical strategy

Strategy / action points:

- Make project proposal for development of DB
  - at least 2 persons: designer & programmer
  - could be part of next CLARIN-D and/or CLARIN-NL
  - TLA is willing to host, participate in, or cooperate with, such a project, provided that funding is guaranteed and that the proposal has the support of a large international community
- Develop a long term plan for the maintenance
  - Perhaps as part of long-term-solution for CLARIN
  - TLA could host and maintain the database, or support the maintenance of it at some other place – funding?

TOPICS

1. Introduction: The FROLIC project proposal
2. Organizational & administrative strategy
3. Technical strategy
4. Content-related FROLIC proposals

4) Content-related proposals

- The sub-standards 1 and 2 are largely fixed and will only very exceptionally need revision
- Also substandard 3 is by now widely used and should be held as constant as possible
- ISO 639-4: specification of procedures & criteria
- To be worked out by a task force in the course of a project, then proposed to a group of experts, possibly those that will be part of the SAB
- The ***criteria*** for entities to be included (or not) in ISO 639-3 have to be general & very explicit

4) Content-related proposals

- Two options for criteria for ISO 639-3:
  - Linguistic (main criterion: mutual intelligibility)
  - Socio-political (splitting & joining ling. langs.)
- Necessary to acknowledge that the topology of languages can be more complex than the usual family — language — dialect distinction
- Already now “macrolanguages” in ISO 639-3
- I am not familiar with the current ISO 639-4
- It has to be practical and globally applicable
- Goal: easy identification of LR (also spoken)

4) Content-related proposals

- ISO 639-3 is recognizing some social-political lgs.
- More comprehensive terminologies may be needed
- For example T. Kaufmann's (1990) proposals:
  - ***Families — languages — dialects***: paradigmatic and most common case
  - ***Language areas*** and ***emergent languages:*** clear boundaries but high intelligibility
  - Some languages are ***dialect chains*** (serial intell.)
  - ***Language complexes*** with ***virtual languages:*** dialect chains with subsets functioning as languages
- Temporal variation complicates things further

4) Content-related proposals

FROLIC-approach to ISO 639-5:

- Acknowledge that there are concurrent proposals for language families and their inner structure
- Accept these proposals side by side
- Characterize the language families with regard to
  (a) acceptedness / disputedness, and/or
  (b) strength of evidence (preferable)
- Distinguish between language families and hypothetical proto-languages for the families
- Relations: “belongs to” / “is derived from”, etc.

4) Content-related proposals

FROLIC-approach to ISO 639-6:

- Current proposals are seen as not feasible
- Perhaps to be replaced or superseded (...-7?)
- Language internal variation is multi-dimensional
- Some variation dimensions are universal & finite
- These can be added as separate tags / properties
- Others are highly language specific
- Entities with proper names should receive a code
- Again, relations of different kind are crucial: “belongs to” / “is derived from” & more specific

4) Content-related proposals

Dimensions of linguistic variation:

- Space (dialects, over-regional standard varieties)
- Time (epochs, periods, stages)
- Social groups (sociolects of several different types)
- Medium (oral, written, signed, whistled, drummed...)
- Situation (registers of different formality)
- Individual ("idiolects", better: "personal varieties")
- (Possibly) proficiency (for learners varieties of different stages, motherese and similar)

Spatial & temporal varieties can be complex themselves

4) Content-related proposals

Strategy / action points:

- Research project for proposing criteria and overall design of ISO 639 -3, -5, -6
- FROLIC is one such project
- Blueprint for (improved) ISO 639-4?
- Once in place, a large international research project should try to cover most regions/families
- Not alone in CLARIN, but global collaboration
- Improving ISO 639 -3, -5, -6, populating DB
- To be approved by the Scientific Advisory Board

4) Content-related proposals

Strategy / action points:

- Points of departure are existing resources:
  - ISO 639-3 / Ethnologue
  - Multitree
  - Glottolog
  - Linguasphere catalog
  - ELcat http://www.endangeredlanguages.com/
  - TEI framework for linguistic variation
- Indicate (a) acceptedness / disputedness, and/or (b) strength of evidence, also for relations

CONCLUSION

- The situation is complex in many cases, and unknown as of yet in many others
- But in principle a pragmatic account is feasible
- With collaborative, pragmatic compromise spirit
- Only possible if strong egos and opinions are controlled, and fights are avoided at all costs
- Else the ISO LR and technology community will move on without us, with horrible consequences for science
- We have to act as a global movement

Towards a new improved setting for the ISO639 standards: the FROLIC approach

Workshop on Language Identifying Codes
Newcastle, 2013-02-09

SEBASTIAN DRUDE
The Language Archive – Max Planck Institute for Psycholinguistics