STANDARDS FOR LANGUAGE CODES: DEVELOPING ISO 639

David Dalby [α], Lee Gillam [β], Christopher Cox [γ], Debbie Garside [δ]

[α] Linguasphere Observatory / Observatoire Linguistique, Hebron, Wales SA34 0XT: dalby@linguasphere.com
[β] Department of Computing, University of Surrey, Guildford GU2 7XH: l.gillam@surrey.ac.uk
[γ] British Standards Institution (BSI), London W4 4AL: christopher.cox@bsi-global.com
[δ] ICT Marketing Ltd, Haverfordwest, Wales SA61 1BW: md@ictmarketing.co.uk

Abstract

The international community, including the International Organization for Standardization (ISO), is currently seeking more granular systems of language identifiers than the widely used tags of ISO 639 parts 1 and 2. There is growing need for the more precise identification and annotation of language-based resources. This paper presents a key response to this need, in which the *Linguasphere Register of the World's Languages and Speech Communities* would provide the referential framework for a future part 6 of ISO 639. The paper discusses the proposed evolution of ISO 639-6 and its relationship to other parts of ISO 639, including its relevance to the definition of meta-data categories in ISO 12620.

INTRODUCTION

Specification of languages is a key parameter in organising documentation, translation and interpretation, and in the identification, assessment and development of social communities. The design and implementation of a comprehensive yet easy-to-use system of language identifiers is essential for supporting historic, existing and future needs for all forms of electronic communication and document distribution.

Accurate and highly granular identification and tagging of languages will lead also to the precise observation and recording of language use and variation. The ways in which local communities within single languages and across linguistic and geographic boundaries refine, adapt, modify and manipulate language provides a means for identifying and locating languages by multiple criteria. Through the observation of our common use of language, we may now approach the possibility of classifying and coding ourselves, and of being able to document and manage our global linguistic resources within the natural continuum of human communication.

The international community has standardised language identifiers within the ubiquitous ISO 639: "Code for the Representation of the Names of Languages". The current parts of this code provide means by which to identify languages based on 2-letter and 3-letter tags, alongside French and English forms for the names represented. The conventions of this standard have been used and adapted elsewhere, for example, by the Internet Engineering Task Force (IETF), which includes also non-ISO codes.

The language identifiers of ISO 639 are complemented by the alpha-2 country identifiers of ISO 3166[1], such as *en-US* for American English or *fr-FR* for French as standardised in France[2]. Here again, the emphasis is on standardised *written* languages, which (unlike spoken forms) may be determined by national authorities and defined in terms of political boundaries. In this context, so-called "standard" or "correct" forms of speech are best considered as by-products of written languages, modelled on the reading of standardised texts and encouraged by national teaching.

The theoretical and practical limitations of the current ISO 639 have been presented, by governmental agencies and others, to the appropriate ISO sub-committee (TC37/SC2). Discussion regarding these limitations has led to the proposed extension of ISO 639 from a 2-part to a 6-part standard. The additional parts aim at the identification of all languages and "collections" of languages, including the mapping of tags among individual sets of identifiers.

DESIGN PHILOSOPHY

The design and implementation of an efficient and comprehensive system of language identifiers, on foundations already laid within ISO 639, involves a responsibility towards future users and a commitment to supporting legacy data also. It is important therefore to step back from the ongoing formulation of the standard and its necessarily intricate and formalised procedures, in order to reflect on what is to be provided, by what means, and most importantly, for what purposes. A system for language codes necessarily requires a theoretical base. We therefore propose the following principles:

1. all spoken, signed and written languages should be recognised and treated as parts of a *continuum of global linguistic communication*;
2. composite codes representing the potentially variable *classification* of languages and language varieties should not be used as identifying tags.

This theory implies that languages (including varieties and collections of languages) should be located in terms of their immediate geographic and linguistic environment, and be identified and tagged uniquely and individually. A result of this theory is that languages are not identified or tagged in terms of their actual or assumed interrelationships. In principle, this will obviate debate on a succession of current "issues", dealing with often impossible questions of how to define individual languages, language varieties and collections of languages.

Any framework envisaged for the future identification and tagging of languages must go beyond the

[1] Both ISO 639 and ISO 3166 tags already form an essential component of the World Wide Web Consortium's eXtensible Markup Language (XML), where the *xml:lang* attribute takes values constructed from at least these two standards as identified in the Internet Engineering Task Force *Request for Comments 3066* (IETF RFC3066).

[2] The capitalisation of country identifiers is non-obligatory, but provides a visual distinction from language identifiers.

requirements and practices of the late 20th century, which were centred on the needs to catalogue and to translate the written word. Such a framework must be simple, precise and elegant, and able to accommodate vast shifts of scale - from humankind's ever-changing communication needs to the future applications of nanotechnology and changes in the political spectrum.

The expected complexity of future needs and applications requires a comprehensive standard (or standards) of linguistic identification and tagging. Such a standard should be as detailed and "granular" as possible, and needs to provide a practical, usable solution to language identification. Debate about the relative status of the items identified can continue, but should not affect the allocation of identifiers to those items.

PRESENT AND FUTURE NEEDS

The increasing mobility and dissemination of language communities has been paralleled by the electronic transformation of the spoken word into the principal medium of worldwide communication and instant documentation. The ability to record, digitise, catalogue and in some cases to automatically transcribe, the spoken word provides an initial set of challenges for computer and mobile technologies. Additionally, the storage of annotated video, for example through automatic transcription, provides for so-called “Grand Challenges” such as “Memories for Life” (Fitzgibbon and Reiter, 2003)

The observation and understanding of linguistic phenomena requires the transparent, accurate and unambiguous identification of every spoken, written and sign language, including each component variety, community and recorded corpus, from the most globalised to the most localised.

This need for a coherent system of linguistic identification will continue to increase as further commoditisation of electronic communications and speech applications occurs: video capture and personal organizers are now commonplace on upper range mobile telephones. As computer devices become faster and cheaper, and are supplied to communities of all sizes around the globe in their own languages, the need for such a system will expand. As needs for communication among communities increase, demands for multilingual translation and interpretation, including subtitling and dubbing, will likewise expand.

The increasing use of XML-based business and administrative communication has already highlighted the need for an extended and structured system of language identification and documentation. As the XML community extends into business communication, and as ISO Committee TC37[3] widens its scope to include resources in speech, so the XML-based implementation of a standardised system of unambiguous language reference becomes essential.

At the same time, the growing multilingual market for electronic communication and universal education will require standardised documentation and coding of languages.

As the multilingual character of megacities develops and changes, there is likewise an urgent need for a system of linguistic and ethnic identification and documentation. Such a system needs to treat languages, not merely as independent information objects, but as integral parts of a worldwide social network of multilingual communication.

LIMITS OF EXISTING SYSTEMS

The original alpha-2 language tags of ISO 639, developed since 1967[4], are an already familiar part of the ICT landscape. These convenient tags for major languages, such as *en* and *fr* referred to above, have a permanent role to play in the standardised identification of the most widely written languages of the world.

Although, in theory, these alpha-2 tags represent the totality of each language, including all its spoken varieties, the tags are used in practice primarily to identify literary and standardised written languages. This is reflected in the criteria for the inclusion of any language in the corpus of ISO 639-1, which is limited to languages for which there is "a significant body of existing documents" and "a number of existing terminologies in various subject fields", in printed or electronic form.[5]

The alpha-2 language tags of ISO 639 comprise the first of two layers in the late 20th century development of that standard. Since 1988, they have been supplemented and to some extent supplanted by the alpha-3 tags of ISO 639-2[6], of which ISO 639-1 is now considered a subset. In contrast to the first layer, with a “terminological” basis, this second layer of ISO 639 was derived from the alpha-3 tags originally employed in North America to identify languages as information objects, not only for bibliographic cataloguing purposes by MARC[7] but also for information interchange by Z39.50[8] (a client/server-based protocol for searching and retrieving information from remote databases).

The second layer of ISO 639, as currently used to identify a restricted range of documented languages by use of alpha-3 tags, is likely to be extended in the near future through unification, or mapping, of ISO 639-2 tags with those used by the Summer Institute of Linguistics (SIL)[9] in the *Ethnologue: Languages of the World* (14th ed., Grimes 2000a, 2000b). It has been proposed by SIL

[3] ISO TC37 has recently formed a fourth subcommittee (SC4) to focus on language resources, for which the further refinement of ISO 639 will be crucial.

[4] As ISO R/639 until 1988, as ISO 639 until 1998, and subsequently as ISO 639-1.

[5] Although the number of speakers of any language is another "consideration" for inclusion in ISO 639-1, some literary languages included are extinct and have no natural speakers.

[6] including alternation between tags based on "own" names (ISO 639-2/T "terminological") or on names in English (ISO 639-2/B "bibliographic"), e.g. [fra] or [fre] for Français (French).

[7] i.e. "**MA**chine-**R**eadable **C**ataloging", a standard format for representing information in a catalogue record in machine-readable form. Since 1987, MARC has been progressively adopted by the wider information community as a convenient way of storing and exchanging bibliographic records.

[8] "Z39.50" covers the joint standards of ISO 23950 and ANSI/NISO Z39.50.

[9] SIL International (formally Summer Institute of Linguistics) is a sister organisation of Wycliffe Bible Translators, benefiting from the collective research and data provided by 5300 missionary linguists, working in more than 1000 languages.

that this work should provide the basis for ISO 639-3[10], alongside a further proposal for ISO 639-5[11], designed to accommodate the development of alpha-3 tags to identify also "language families" and "language groups".

A further part of the ISO 639 standard, ISO 639-4[12], is currently being developed as a statement of guidelines and principles for the future development of language codes in general. It is hoped that the present paper may serve as a useful contribution to the discussion and considered evolution of this essential part of the overall standard.

Requirements for a comprehensive system of language identifiers in the 21st century need not be limited to the 20th century convention of 2-letter and 3-letter attempts to provide mnemonics for commonly used language names. The number of entities to be identified in any future system, including all known languages past and present, and all varieties and collections of languages, is already in excess of 25,000[13]. By contrast, ISO 639-1 contains less than 150 of a usable 676 alpha-2 combinations, while ISO 639-2 and the proposed 639-3 (SIL) codes contain less than. 400 and 7,200, respectively, of a total usable 17,576 alpha-3 combinations.

A system of alpha-4 codes has therefore been essential, allowing the selection of around 25,000 identifiers with room for expansion within a total limit of over 450,000 potential combinations.

PROPOSED USE OF ALPHA-4 TAGS

Following publication of the *Linguasphere Register of the World's Languages and Speech Communities* (Dalby 2000a, 2000b), Technical Committee TS/1[14] of the British Standards Institution (BSI) extended an invitation to the Linguasphere Observatory[15] in Wales to propose a standardised system for tagging all the world's languages and for identifying their inter-relationships, present and past. The first draft of this system, Linguasphere System 639 (LS 639) is now complete, and we hope that it will be possible to make the index of this system available from mid-2004 (see http://www.linguasphere.com/)

LS 639 contains an index of over 70,000 tagged language names, including "dialects" and language "groups". Each tag gives access to information on and in each relevant language, and its components, and enables the information to be viewed in the context of the language's wider relationships. LS 639 is cross-referenced to ISO 639-1, 639-2 and the proposed 639-3, and provides and defines a series of over 25,000 unique alpha-4 tags. These are intended to cover every known language, written, spoken, and signed, either modern and/or recorded from the past, as well as a growing catalogue of the component dialects and speech communities within individual languages. The same form of alpha-4 tag is applied also to groupings of two or more languages, whether established or hypothetical.

The totality of Indo-European languages, for example, is identified by the same form of alpha-4 tag, in this case **/ineu/** and the local form of the Welsh language spoken around the Preseli hills of west Wales, is identified by **/prsl/**. Between these two extremes, alpha-4 tags are used also to identify the set of Celtic languages within Indo-European, **/celt/**, the net of "Brythonic" or Britannic languages within Celtic, **/brtn/**, and the "Welsh" or Cymraeg language itself, **/cymr/**.

The application of fixed alpha-4 tags to all levels of linguistic identification has the following advantages of simplicity, precision, transparency and flexibility:

1. With over 450,000 potential combinations, LS 639 is able to represent the actual scale of complexity of spoken languages around the world.[16]
2. The mnemonic form of LS 639 tags favours human readability alongside essential machine readability.[17]
3. High granularity gives LS 639 a refined power of definition, allowing "languages" to be identified in terms of their components rather than the reverse.[18]
4. LS 639 supports the parallel use of ISO 639-2, with its proposed extensions (639-3 and 639-5), since each alpha-3 tag will be precisely definable in terms of its alpha-4 equivalents, covering its components and wider linguistic context.[19]
5. The correlation of LS 639 tags with all other forms of language identifiers will support all legacy databases with fixed 2- or 3- character fields for language identifiers.
6. The use of alpha-4 tags at all levels will facilitate, whenever required, the future redefinition of any "language" as a "variety" of a wider language, or as a "collection" of two or more languages, without changing its LS 639 tag.
7. Each fixed alpha-4 tag is located by reference to its coded and potentially variable place on the LS 639 relationship scale.[20]

CLASSIFICATION AS METADATA

The final pair of attributes presented above have enabled LS 639 to solve a hitherto intractable problem. The classification of linguistic relationships provides an obvious framework for organising data on natural languages. Yet how can such a framework be protected from the inevitable upheavals caused by any reassessment

[10] Proposal: ISO/CD 639-3 dated 2003-08-29,for "Codes for the representation of names of languages - Part 3: Alpha-3 codes for the comprehensive coverage of languages".

[11] Proposal: ISO/WD.1 639-5 dated 2003-11-26,for "Codes for the representation of names of languages - Part 5: Alpha-3 codes for language families and groups".

[12] Proposal: ISO/WD.1 639-4 dated 2003-11-28,for "Codes for the representation of names of languages - Part 4: Implementation guidelines and general principles for language coding".

[13] The *Linguasphere Register* exceeded the total of 25,000 entities in the year 2000, and the 2nd edition in 2005 will mark the first stage of its continuous refinement and expansion.

[14] UK shadow committee to ISO Technical Committee TC37, concerned with "Terminology and other language resources".

[15] created in Quebec in 1983 and subsequently established in Normandy as l'Observatoire linguistique (an "association 1901" under French law), with its research centre in Wales since 1995.

[16] In contrast to alpha-3 tags, limited to just over 17,500 options.

[17] Machines require no mnemonics, but speakers are likely to prefer the meaningful tagging of their languages.

[18] In contrast to alpha-3 tags, which must depend on the *a priori* definition of individual "languages": see Constable *loc cit*.

[19] In this context, the Linguasphere Observatory welcomes close consultation with ISO TC37 and SIL/*Ethnologue*.

[20] See *Linguasphere Register*, vol.1, pp.58-70.

of linguistic relationships[21]? One thinks of the way in which books on African languages, for example, needed to be reclassified in the mid-20th century to cater for major changes in their classification.[22]

This problem has been catered for in LS 639 by treating the comprehensive *identification of inter-relationships* among languages as a fundamental category of metadata, attached to but not determining the alpha-4 identifiers of individual languages or varieties of language. A continually updatable *roadmap of the linguasphere* may consequently serve as a logical supplement to – but not necessarily a part of - the proposed expanded structure of ISO 639.

Alongside the provision of fixed identifiers for all components of the linguasphere, from dialects and speech communities upwards, there needs to be a separate allocation of modifiable codes to all components, according to their current place in a scheme of universal geolinguistic relationships [23]. This scheme is an essential part of the meta-data attached to the LS 639 identifiers, and will be presented for discussion within the context of the 2004 LREC workshop on the Registry of Linguistic Data Categories.

DISCUSSION

The system of LS 639 identifiers, proposed as a basis for ISO 639-6, has been designed to provide a route for the further development of ISO 639, as represented by the existing alpha-2 and alpha-3 tags. This expanded system will provide a necessary "road-map" for the adoption of, and optional migration to, its more extensive set of alpha-4 identifiers. The dissemination and use of such a system will be important in the fields of business, government, education, social research and the media.

Assisting international consortia by the introduction and use of the LS 639 system will be a valuable scientific contribution from Europe. The system will also include the geographical mapping of alpha-4 coded items. Some of this work has already been undertaken among members of the Linguasphere network, including cartography in UK (centred on Africa), in France (centred on the Himalayas) and in Russia (centred on the Caucasus).

The final system, along with revision and development of other parts of ISO 639 should facilitate referential transparency. For example, when we refer to the "English language" it is often unclear whether or not we are referring to the standardised written (and spoken) language. Are we ignoring the minor differences between American, British and other conventions in the standard written language? Alternatively, are we referring collectively to *all* forms of the English language, including every spoken "dialectal" variety in the world and all recorded written forms, past and present?

The increasing freedom of communication in whatever form of spoken or written language individuals choose to use (including, for example, new forms of abbreviated spelling commonplace in text communication on mobile phones), makes it more important that the ISO 639 standard should make provision for specifying each standard language as opposed to identifying the fuzzy totality of each major spoken and written language. The LS 639 proposal provides the means for satisfying this requirement.

The proposed expansion and refinement of ISO 639 coincides with proposed and ongoing work regarding the development of metadata registries for language resources (by sub-committee TC37/SC4). Metadata registries for language resources, described in accordance with ISO 11179-3, will enable systems to make direct reference to metadata defined according to standards. In this case, language identifiers will become *keys to meta-data* within these registries. Based on work carried out initially in ISO 12620:1999, which described so-called "Data Categories" found in terminological collections, ISO 12620 is being revised in conformity with ISO 11179-3 to describe the management of data categories, with subsequent parts providing descriptions of validated data categories, for example part 2 for terminological data categories. The parallel development of these (sets of) standards will provide a link between the creation and management of language identifiers and their management and use within software systems via metadata registries, enabling and ensuring interoperability between language resources that may use differing systems of language identifiers, (at the very least). Use of Data Categories for specific types of language resources has been described for terminologies in ISO 16642 (Terminological Markup Framework).

[21] Historical relationships among languages are sometimes described as "genetic". This is misleading in that languages are not independent objects when in close contact within the minds of bilingual speakers, who are the key players in the evolution of the linguasphere.

[22] When major groupings such as "Sudanic" were replaced by new groupings such as "Niger-Congo".

[23] See *Linguasphere Register* (Dalby 2000a), pp.47-74.

References

Dalby, D. (2000a) "Linguasphere Register of the World's Languages and Speech Communities: Volume 1 (Introduction and Index)". Hebron (Wales). ISBN 0 9532919 1 X

Dalby, D. (2000b) "Linguasphere Register of the World's Languages and Speech Communities: Volume 2 (The Register)". Hebron (Wales). ISBN 0 9532919 2 8

Fitzgibbon, A. and Reiter, E. (2003) "'Memories for life': Managing information over a human lifetime". http://www.nesc.ac.uk/esi/events/Grand_Challenges/proposals/Memories.pdf (4 March 2004)

Grimes, B.F. (Ed.) (2000a) "Ethnologue: Volume 1 Languages of the World" 14th Edition 866 pp., ISBN 1-55671-103-4

Grimes, B.F. (Ed.) (2000b) "Ethnologue: Volume 2 Maps and Indexes" 14th Edition 735 pp., ISBN 1-55671-104-2

ISO639-1:1988 "Language Codes – Part 1: Alpha-2 code"

ISO639-2:1998 "Language Codes – Part 2: Alpha-3 code"

ISO3166-1: 1997 "Country codes"

ISO11179-3:1994 "Information technology – Specification and standardization of data elements. Part 3: Basic attributes of data elements".

ISO12620:1999 "Computer Applications in Terminology – Data categories"

ISO16642:2003 "Computer Applications in Terminology – Terminological markup framework (TMF)"