L2/16-132

Review and comments on TC37/SC2 NWIP: Identification and description of language varieties

Peter Constable, 2016-5-5

ISO TC 37/SC2 is balloting a new work item proposal for a new international standard, tentatively “Identification and description of language varieties”. I have quickly reviewed this and have the following initial reactions:

This proposed standard would create a framework for description of language varieties by defining a set of dimensions by which language can vary. The targets of description are individual language resources, or instances of language usage — utterances, for example. Eight dimensions are defined, and are assumed to be comprehensive and complete. These include space (geographic location), time, social group, modality, and others. The proposed standard would not specify any enumeration of values for these dimensions, however. Nor would it attempt to create a catalog of all varieties of all languages. It suggests the creation of a registry for attribute values as a logical next step.

This proposed standard is reminiscent of ISO 639-6, which was an attempt to create identifiers for language varieties. (ISO 639-6 has since been withdrawn as an international standard.) However, there are important differences. First, whereas 639-6 attempted to create a comprehensive catalog of language varieties, this proposed standard does not. Secondly, whereas 639-6 defined identifiers for each variety, that does not appear to be an intended goal for this proposed standard. In addition, and most significantly, whereas 639-6 did not include any operational criteria or model by which varieties could be determined, the main intent of this proposed standard is to provide a framework for characterizing varieties. It still would not specify how to assign particular attribute values along any of the dimensions, so there is still a largely ad-hoc aspect to defining language varieties as there as in 639-6. But, at least, it defines a set of dimensions to be used in defining varieties, which 639-6 did not do, and it explicitly does not aim to define a comprehensive enumeration of varieties, and 639-6 did — which was misguided, in my opinion.

Clause 3.3 of the working draft that supplements the NWIP presents terminology pertaining to various factors that may be involved in each of the different dimensions. For example, in relation to the space dimension, “space”, “dialect” and “standard variety” are defined; in relation to the social group dimension, “social group”, “sociolect”, “technolect” “domain” and “subject” are defined. What is not explained in clause 3 or elsewhere of the working draft is how these different factors inter-relate in assigning a characterization for a variety in regard to a given dimension.

The proposed standard anticipates use in applications involving language resources. It does not define any specific data representation formats. It assumes that its conceptual framework would be incorporated into different data representation frameworks in

different ways appropriate to each formal framework. As example metadata frameworks, it cites the Open Language Archive Community Metadata standard (http://www.language-archives.org/NOTE/usage.html) and the Component Metadata Infrastructure best-practice guide (http://www.clarin.eu/content/cmdi-best-practice-guide). It asserts its conceptual framework to be an appropriate extension for such existing metadata frameworks. One possibility mentioned in particular is that a BCP 47 extension could be defined using the conceptual framework of this proposed standard as its basis.

The working draft asserts its framework as suitable for meeting needs in the domain of language technologies. It cites as a selling point that it "complies to the 'Recommendation on software and content development principles 2010'," a statement that was evidently formulated at the 2010 International Conference on Computers Helping People with Special Needs (ICCHP).

Of potential interest for the Unicode Consortium, these recommendations appear to focus on software development, with a succinct (perhaps a bit over-simplified) summary being, "Follow i18n best practices." While ICCHP is not a significant driver for i18n best practices within the software industry generally (at least, in North America), it may be the case that it has a more prominent role in the assistive technology sector. (I'm not familiar enough with that sector to evaluate that.) Within the assistive technology sector, language technologies such as text-to-speech are one important segment. Historically, the Unicode Consortium has not had much direct engagement with the assistive technology sector: the likes of Nuance, Meridian One, Acapela Group, or other developers of assistive, language technologies have not been members or participated actively in our technical committees. My point in all this is that it's not clear to me to what extent the proposed standard would meet actual needs in the area of assistive, language technologies. It doesn't seem to meet any key needs in areas that the Consortium generally deals with, but our needs in relation to identification and characterization of language varieties may not be representative of all technology sectors that deal with linguistic data.

Assessment of this new work item proposal:

It seems unlikely that this proposed standard would be used in any Unicode projects, at least in the near term. But it's also not made clear in the proposal what specific sectors might be expected to adopt it. It seems like the conceptual framework could be useful for linguistic research, language documentation and housing of large language corpora. But there is no indication that this framework is currently implemented and in use within any user community, and it's not clear if there are any particular user communities endorsing this particular framework. For this reason, I question whether adoption as an *international standard* is appropriate at this time. I'm reminded of ISO 639-6, which was made an international standard under similar circumstances: a solution without a clear, significant user community waiting to adopt it. Developing that as an international standard was a failed business plan. Now, I believe that this framework may have much better potential for success than 639-6 did. Even so, pursuing an international standard from the outset seems premature.

Besides the uncertainty about likely, actual usage, it's not clear that the proposed framework, in its current stage of development, has characteristics appropriate for an international standard. In particular, there is no statement regarding conformance. Moreover, there is no clear way of defining conformance beyond a very minimal level: it could require that conforming implementations characterize language resources in terms of the specified dimensions of variation, but without any attribute values defined, there would be no way to evaluate if actual instances of metadata actually fit the model. For example, a given resource might have the attribution, [register = "Brillig"], but without any conventional or specified definition for "Brillig", there would be no way to evaluate if this could, in fact, be considered a characterization in the dimension of register. Furthermore, any implementation could incorporate ad hoc extensions, using arbitrary dimensions and values, and this standard would make no judgment for or against that practice. Thus, there is no means of enforcement. Any claim to conformance would be effectively meaningless.

This is not to say that the framework has no potential for eventual adoption as an international standard. A framework of this nature has far better potential, in my opinion, to be a useful basis in an international standard than the framework underlying ISO 639-6 ever had. But I think it needs to develop and mature further, and needs to acquire a clear customer base before adoption as an international standard. In particular, the suggestion of creating a registry for attribute values in each of the dimensions seems essential in order to enable real interoperability of actual instances of metadata based on the framework. Along the way, I would expect pre-standardization implementations might serve to refine the framework, helping to identify potential quirks in the model that would ideally be resolved prior to reaching a status at which stability of the framework begins to be an important consideration for interoperability.

With these things in mind, I would offer the following comments to TC37/SC2:

- At this stage, this project should be developed as a Technical Report, or at most, a Technical Specification, and not an International Standard. The framework requires more development, and a more clearly-identified adopting customer base before it can actually succeed as an International Standard. I would also encourage TC37/SC2 to have a longer-term plan that includes further development beyond a Technical Report.

- The working draft that accompanies the NWIP states, "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." As noted, such a registry is an essential pre-requisite for any mechanism that can provide interoperability of metadata. Therefore, a minimal requirement to succeed as an international standard should be the establishment of such a registry or at least a specification of requirements for any applications that might create such a registry.

- The working draft cites the IETF BCP 47 specification and its provision for defining extensions, suggesting the possibility of a BCP 47 extension based on the proposed framework. Such an extension would need to specify or make normative reference to a

registry of language-variation elements. Specification of such an extension to BCP 47 might be a very useful way to establish not only a registry of semantic language-variation attribute categories, but also a formalized metadata protocol for interchange of language variation descriptions / identifiers. This may be a useful direction for development beyond a Technical Report that TC37/SC2 may want to consider.

- The NWIP and accompanying WD provides a model for describing language variations in terms of eight dimensions of variation. It is not clear, however, to what extent this model has be researched and exercised in actual usage in order to evaluate its sufficiency and fit as a best-practice model for the intended purposes. If such research has been conducted or if there has been prior usage, references to such work should be included in a NWIP. If not, that would underscore that this proposal, however promising, is premature for consideration as an international standard.

- The NWIP cites speech technology as one sector for which the proposed metadata framework will be an essential prerequisite to sustainability and interoperability of resources. Yet no information is provided indicating how the proposed standard (or specification) would relate to any metadata frameworks that are in actual usage in the speech technology sector. Nor is there indication of active support from any organizations directly connected with the speech technology industry. The only liaisons suggested are within TC37 itself, whereas, with an emphasis given to speech technologies, one might expect liaison relationships with some external agencies dealing with speech technologies, such as the W3C (with relevant specifications including VoiceXML and Pronunciation Lexicon Specification); and at a minimum, with JTC1/SC35, which has a scope that encompasses speech and assistive technologies. The NWIP should, preferably, provide indication of support for or partnership in the development of a standard or specification from outside TC37.