Language identification and IT
Addressing problems of linguistic diversity on a global scale*

Peter Constable and Gary Simons, SIL International

Abstract

Many processes used within information technology need to be customized to work for specific languages. For this purpose, systems of tags are needed to identify the language in which information is expressed. Various systems exist and are commonly used, but all of them cover only a minor portion of languages used in the world today, and technologies are being applied to an increasingly diverse range of languages that go well beyond those already covered by these systems. Furthermore, there are several other problems that limit these systems in their ability to cope with these expanding needs. This paper examines five specific problem areas in existing tagging systems for language identification, and proposes a particular solution that covers all the world's languages while addressing all five problems.

1. Introduction

The information technology (IT) industry has been driven in recent years to address problems of multilingualism and internationalization. This has been driven to a significant extent by the growth of the Internet. Rapidly increasing economic development throughout the world, together with the growth of the 'Net, has actually resulted in a significant increase in the number of languages that technologies need to support. In many parts of the world, speakers of previously "unknown" languages (that is, unknown to speakers of "major" languages) are beginning to make their mark on the World Wide Web, and are using their own languages to do so.

Even apart from the Internet, communities of speakers of lesser-known languages are using technology to pursue linguistic development of their communities through literacy, literature development and other means. In addition, researchers such as linguists and anthropologists, development and relief organizations, and governments are pursuing interests involving thousands of different linguistic and ethnic communities around the world. In this work, they are seeking to make use of current information technologies, such as Unicode and XML.

---

* This is a revised version of a paper that was presented at the 17th International Unicode Conference in San José, California in September, 2000, and which appears in the conference proceedings. For the most part, the changes include some additional examples that were used in the actual presentation at that conference, and some minor additions dealing with some of the issues that were raised in the discussion that followed.

We wish to thank the anonymous reviewers for their helpful comments. Any remaining shortcoming are, of course, our responsibility alone.

This document is available online at http://www.silewp/2000/001/.

For a variety of reasons, discussed briefly in section 2, it is important to tag information objects to identify the language in which information is expressed. There are, however, a variety of issues to be faced in tagging information objects to identify the language. We have identified these key problems:

- Change: Because of the very nature of human languages, and because of the difficulty in obtaining complete and accurate knowledge of languages, it is impossible to create a static categorization of all languages.
- Categorization: Different operational definitions of *language* that can serve different purposes lead to different categorizations of languages that may not agree with one another.
- Inadequate definition: Existing systems of language identifiers do not employ consistent operational definitions and in many cases list objects that are not languages per se.
- Scale: There are on the order of 6,800 languages known to exist in the world today, which is an order of magnitude greater than what existing systems of language identifiers currently cover, and existing systems do not scale well.
- Documentation: Existing systems of language identifiers do not adequately document what language-related category is denoted by a given identifier; in most cases, they provide nothing further than a name, which is inadequate in most cases.

In sections 3–7, we will discuss each of these in turn, examining the nature of the problem and identifying any implications for IT. Then, in sections 8 and 9, we will review what is needed to solve the problems we have outlined, and present a specific proposal that we believe would provide significant advances in solving these problems.

In this paper, we will consider some existing systems for language identification in use within IT today. These will focus primarily on three particular systems, because of their widespread use and importance within IT: Win32 LANGIDs [6] (used in Microsoft Windows); RFC 1766 [1] (used in XML [7]); and ISO 639(-x) [3, 4] (used in a variety of IT systems and as a basis for other IT standards).

2. The need for language identifiers

Humans usually don't have any need for language identifiers: if they hear somebody speaking or read something and the language is one that they know, then they recognize it immediately and are able to process it. Of course, this may not be true 100% of the time; for example, on encountering the written word *chat*, one doesn't know if it's English (for 'talk informally') or French (for 'cat') or some other language. Information is usually provided in a context that eliminates such ambiguities, however. If, on the other hand, we don't recognize the language, the processing stops. In this case, knowing the name of the language would not help: we still couldn't process it. In all of this, we don't need a label that identifies the language for us. Our brains do need to identify the language before we can process it, but we almost never need a label to do that.

Software processes are like human brains in that they do need to identify the language of linguistic information before certain kinds of processing can be performed on the information. But unlike human brains, identification cannot, in general, be based upon an analysis of the information itself. While it is true that advances in language recognition have been made (so, for example, some commercial business applications are beginning to employ language recognition techniques), these are still very limited and available for only a very small number of languages. Even if such technology could reach a point in the future where it is highly reliable and available for every human language, it would simply not be practical to depend upon this mechanism in every situation. For example, it isn't practical when searching for documents in a particular language on the Web, and it would not work in multilingual documents that have short runs of text in a given language.

There are a wide variety of processes for which it is necessary to identify the specific language beforehand. Language-based indexing and searching are fairly obvious examples, as is semantic interpretation. But there are a number of others: spell-checking, sorting, syllabification and hyphenation, morphological and syntactic parsing, fuzzy string searches and comparisons, speech recognition, speech synthesis, semantic associations, thesaurus lookups, and potentially many others.[1] In this paper, we will refer to all such types of language-specific processes collectively as *linguistic processes* or *linguistic processing*.

It is important to distinguish between identifying the language that is used in an information object and languages that the information may describe or refer to. The latter is relevant, for example, in bibliographic cataloging to identify languages for subject headings. The identification needs of linguistic processes and of subject indices are, in general, very different. For a subject heading, a generic linguistic classification such as "Bantu languages" may serve a useful purpose. In these cases, the search engine requires only a label and does not need to know any specific language-related information.

For linguistic processing, however, such generic classifications are not useful. For example, if all that is known about a document is that the language it is written in is a Bantu language, it is not possible to spell-check that document. In order to perform such linguistic processing, the exact speech variety (language or dialect) must be known.

The main point to be made here is that textual information objects (or, indeed, any digital representations of language use) need to be tagged to identify the exact language of the object if there is a potential need to support subsequent linguistic processing of various kinds on that information. In that users may wish to work with information in potentially any language, it is necessary to have systems for language identification that provide identifiers for all languages.

3. The nature and scale of linguistic diversity

In today's software market, an application that is enabled for 50 languages is considered to be very multilingual. In contrast, by one well-researched estimation, there are some 6,800 living languages spoken in the world today. The following tables show statistics for language distribution by region of the world, and for countries having the largest number of living languages:

Region	# of languages
Africa	2062
Americas	1020
Asia	2202
Europe	237
Pacific	1312

Table 1. Number of languages by region [2].

[1] It will be noted that, in many of these cases involving textual information, it is also necessary to identify the writing system or the system of orthographic conventions in use as well as the language. It turns out that many processes actually require identification of *para-linguistic* entities—notions that combine language per se with notions that may be related to but are different from language, such as orthography, sort order or location. There are several different para-linguistic notions that may be required by IT processes, and there is a need to develop an adequate conceptual model of the various para-linguistic notions that are needed in IT and of their interrelationships. But the notion of language proper is assumed in the definition of all of these other notions, and so an adequate model for addressing language identification must first be established. Thus, we have chosen to limit the current discussion to identification of languages.

Country	# of languages
Papua New Guinea	823
Indonesia	726
Nigeria	505
India	387
Mexico	288
Cameroon	279
Australia	235
Congo (DRC, formerly Zaire)	218
China (PRC)	201
Brazil	192
USA	176
Philippines	169

Table 2. Countries with more than 150 living languages [2].

Some may find the numbers surprising and wonder how there can be so many languages in the world. We will briefly explore some of the causes behind this diversity, and then consider consequences for systems of language identification.

3.1 Causes of linguistic diversity

One unquestioned universal of human language is change: no two people use language in exactly the same way, and even a given individual's use of language changes over time.

With the passage of time, a given linguistic community can experience a wide variety of linguistic changes: pronunciation alters, morphological paradigms become simplified, new lexical items are introduced while others are abandoned, idioms are coined and then become fossilized, semantic ranges shift, syntactic constructions are modified, societal attitudes toward language use are transformed. The net effect can be that, after several generations, the language spoken by a language community may be called by the same name as the language used by their ancestors, yet be significantly different from it.

Not surprisingly, social need creates a tendency for a group of speakers living in close proximity and having regular social interaction to maintain a common form of speech, undergoing linguistic changes together. On the other hand, if a community splits apart and the parts become physically or socially isolated from one another, internal divergence will occur in which the forms of speech of the separate communities drift apart as each undergoes different linguistic changes. This change is not always gradual; for example, there may be wholesale borrowing as one community comes in contact with another. More generally, when communities speaking different languages come into close contact with one another, each may influence the way the other speaks. When a speech variety changes in this way by convergence with an external variety, this has the effect of causing it to diverge further from the varieties it used to be like.

When a given speech variety has undergone a relatively limited degree of internal divergence, dialects are born. Eventually, the different communities may no longer be able to understand one another, and distinct languages are born.

Over the course of thousands of years of human history, people groups have migrated and settled into nearly every corner of the earth. As this has happened, the processes of external convergence and internal divergence have been in continuous operation. The result has been a complex web of multilingual diversity.

For example, it was noted in table 2 that more than 800 languages are spoken within the country of Papua New Guinea. This is among a population of less than 5 million living in an area only slightly larger than the state of California. Within the country, rugged mountains have formed effective barriers that have kept

groups apart on a macro level. On a more local level, inter-group rivalries and warfare have historically kept relatively small communities isolated from one another.

It is even more fascinating to note that a country the size of Vanuatu, with less land mass than the state of Connecticut and a population of under 200,000, has over 100 living languages. That gives an average population per language of under 2,000 speakers. This country consists of a chain of some 80 islands, and in this case, it is the Coral Sea that forms a physical barrier that has kept groups isolated from one another.

The majority of the world's languages are spoken by language communities with relatively small populations, often in remote locations, such as the mountain valleys of Papua New Guinea or the islands of Vanuatu. These facts do not mean that these languages are irrelevant for IT, however. All of these languages are collectively of interest at least to linguists and anthropologists, to governments, and to various development agencies. They are also of interest to the speakers of the languages themselves, who are increasingly attempting to establish a bridge between their languages and the realm of IT. Every language is significant to some portion of the world's population, who will need a way to identify it when they encode it in a data stream.

Because most languages are spoken by small populations living in remote locations, and because of the sheer number of languages involved, determining the complete inventory of languages spoken in the world is an incredibly difficult research task. The information we have presented here is based on the *Ethnologue* [2], a catalog of the world's languages that has been compiled over a period of more than fifty years based on information gathered from many published sources and from a large number of field linguists. Even after all of that time and effort, new languages continue to be identified as better information becomes available. Sometimes this is due to the discovery of a previously unknown speech variety, but most often it results from discovering that varieties previously thought to be dialects of the same language are, in fact, distinct enough to be considered different languages. Such improvements in our knowledge are likely to continue for many years to come.

3.2 Implications of diversity and change for language identification

There are two implications to be drawn out of the preceding discussion. First of all, given the extent of linguistic diversity that exists in the world, systems for language identification must have the capacity to accommodate several thousand distinct languages. This applies both to the mechanism employed for tagging objects, and to the procedures for assigning tags. These issues will be discussed further in section 6.

Secondly, given that languages are constantly changing, systems for language identification must be able to accommodate this change. This means that systems cannot be static: it is impossible to create a complete list of living languages that will remain fixed over time. Inevitably, some of the languages will die out, and other languages will divide, resulting in the need for new entries.

One might argue that, given the current world situation, it is more likely that languages will die, and that those that survive will stabilize, with the result that tags for new languages will never be needed. This is unlikely to be the case for all languages since there are so many potential factors that can lead to separate language identities.

In addition, given that our knowledge of the languages spoken in the world is constantly improving, with new languages still being discovered, systems for language identification must also be able to accommodate these changes in what is known about the world's languages. These changes are not unlike the kinds of changes that occur over time in the languages themselves (e.g. a single language splitting to become two or more distinct languages), and for the most part, amount to adding new languages to the inventory. Accordingly, it must be possible to add to an inventory of language identifiers over time and to maintain the best possible information about what each identifier denotes.

4. Problems of linguistic categorization

It was mentioned earlier that, within a given language community, no two people use the language in exactly the same way. A language as used within some language community will have ranges of variation along multiple axes: variation in pronunciation, variation in syntax, variation in lexical items, variations in collocation, and so forth. Often these variations will correspond to geographic distribution, but variation can exist within close geographic proximity. For example, there may be variation that corresponds to socio-economic variations. In terms of geographic distribution, it is possible for related but distinct languages to be spoken in two locations, A and B, but also for there to be a continuum of variation as one moves from A to B with no distinct boundary at any point along the way.

These issues raise some important questions: What do we mean by *language*? How do we define the limits of a language? What is the difference between a *language* and a *dialect*? In any attempt to identify and enumerate languages, one must first adopt some operational definition for what a *language* is.

In section 4.1, we will examine the problems related to selecting an operational definition of language, reaching the conclusion that alternative definitions are possible, according to different perspectives and needs, and may be equally valid. We then consider the consequences for IT that result from this.

4.1 Selecting an operational definition of *language*

Any attempt to categorize all of the world's languages must assume some operational definition of language. There is, however, no single, objective definition for language. In adopting an operational definition of language to be used in categorizing the world's languages, several issues must be considered:

What factors will form the basis of an operational definition of language?

There are many factors that may be considered, such as the following:

- actual linguistic similarity between speech varieties;
- intelligibility, i.e. the ability of speakers to functionally communicate with one another;
- literacy and ability to share a common literature;
- ethnic identities and self-perception of language communities;
- other perceptions and attitudes based on political or social factors;

A definition can also be based on some combination of such factors. Generally, the combination of factors that one chooses will be based upon one's felt needs and purposes.

What relative priority will be given to the multiple factors?

Linguists generally give preference to factors of linguistic similarity and intelligibility because such factors generally suit their purposes best. Others may have a different basis for prioritizing factors, however. Again, prioritization is generally determined on the basis of felt need and purpose.

How will these factors be measured?

Whatever factors are used, measurements based on those factors must be obtained for use in comparing speech varieties. In a small number of simple cases, clear distinctions may be fairly self-evident, but in general this will not be the case, and some form of analysis of the situation will be needed. Quantifiable measures are more conducive to objective analysis, but not all factors necessarily lend themselves to quantifiable measurement in obvious ways. (How, for example, does one quantify ethnic identity and self-perception?) Even when a means of quantifiable measurement seems obvious, it may not in fact be easy to implement.

What threshold values will be used in assessing the measurements?

Once measurements are obtained, one has to evaluate them and decide how to apply them to the task of uniting or distinguishing speech varieties. Factors will generally involve a continuous range of variation, and so one must decide at what point in that range to draw the line between joining and splitting.

There are, therefore, many ways to form an operational definition of language, and different definitions can lead to different conclusions. This is particularly true of the choice of factors that are given priority.

When casual observers point to two major languages and say that they are different languages, they may not have any specific operational definition in mind, but often the two speech varieties would be distinguished on the basis of several definitions, so the distinction can be considered to be self-evident. When attempting to categorize all of the world's languages, however, this is rarely the case, and one must give careful consideration to the operational definition that is used.

Even in the case of relatively "major" languages, different operational definitions can lead to different conclusions. For example, Serbo-Croatian has often been considered to be a single language with two writing systems, but there is a recent trend toward considering Serbian and Croatian to be distinct languages. For political and other reasons, separate identities are emerging, and many people, employing an operational definition that gives priority to social or political factors, interpret these developments as an emergence of distinct languages. In contrast, an operational definition that gives priority to intelligibility might lead to the conclusion that the speech varieties in question still comprise a single language.

Since many operational definitions for language are possible, there are a corresponding number of ways to categorize all the world's speech varieties into languages. Thus, there is no possible consensus among all experts. In general, each can be driven by differing needs.

Each one, of course, considers their needs to be valid. Bibliographers and librarians have a valid need for language identifiers that correspond to generic language classifications, such as "Bantu languages", to be used as subject headings. But as mentioned above, these do not serve the needs of linguistic processing.[2]

Within SIL and the *Ethnologue*, we have used an operational definition of language with a primary criterion of mutual non-intelligibility—that is, two people speak different languages when neither can understand the other at a functional level. This definition serves our needs in several regards: it identifies speech varieties that are of broad interest in linguistic research, and it identifies speech varieties that are of greatest interest for purposes of language-related development, such as literacy. In our experience, this definition also generally identifies distinctions that are relevant for most linguistic processes.

To summarize, different categorizations of the world's languages are possible because different operational definitions of language are possible, with different definitions generally corresponding to differing needs. Therefore, no consensus on categorization is, in principle, possible, and any categorization of languages should identify the operational definition used and make every attempt to apply it consistently.

4.2 Language categorization and IT

Since different categorizations based on different operational definitions are possible, this raises questions of how the problem of categorization should be dealt with in the context of IT. Should one particular approach to categorization of languages, based on a particular operational definition, be singled out? Should multiple operational definitions for language be accommodated? If so, how is this managed?

---

[2] One could argue that systems of language identifiers for use in bibliographic subject cataloguing should be distinct from systems for use in linguistic processing. Another might counter, appropriately, that bibliographers need language identifiers not only for subject headings, but also to indicate the language in which a work is expressed so that users can select items based on languages that they understand. We would agree, and contend that the latter need is different from the needs of subject indexing and similar to that which is needed for linguistic processing. A generic classification does not help a user in this situation: if a catalog indicates that an item is in "Quechuan" and this item is retrieved by a speaker of Cuzco Quechua, it would not be useful to them if the item turns out to be in a variety of Ancash Quechua.

Certainly, the worst situation would result if no explicit consideration is given to the issue of operational definitions. This would be the case, for example, if a system has a single namespace for identifiers without any consistency in the types of category identifiers denote and no indication of this for any given identifier.

On the other hand, if a specific operational definition is to be singled out, there must be consensus on what that definition will be. Given that different definitions typically correspond to different user needs, it is not clear whether this would be possible, or even desirable, since different needs can be equally legitimate. This becomes a question of whether there is a single approach to creating a namespace of identifiers that maintains order in terms of operational definitions while also meeting the full variety of needs that IT must meet. It is not at all clear that this would be possible.

Another alternative that solves these problems is for a system to allow multiple namespaces of identifiers, each one using different but clearly-defined operational definitions. This allows for a wide variety of needs while maintaining consistency. The main concerns would be that of managing multiple namespaces, potentially greater difficulty in determining the meaning of a given identifier (although this should not be a problem if handled carefully), and potential difficulty when presenting users with options in deciding which namespace is appropriate in a given application.

5. Existing problems of definition

In the previous section, we discussed the issue of operational definitions of language in general terms, and saw that different categorizations of language can result when different operational definitions of language are used. In this section, we wish to point out that some important, existing systems for language identification do not use consistent notions of what the objects are that they are identifying. In general, it is not, in fact, clear that any particular operational definition has been employed at all. Thus, we have namespaces for "languages" with identifiers for categories that are not directly related to languages at all (e.g. they are *para*-linguistic) or that correspond to language-related categories of different types.

5.1 Systems that identify categories not directly related to language

Win32 (the programming interface for Microsoft Windows) uses a system of language identifiers known as *LANGIDs*. The Win32 LANGIDs contain numerous examples of categories that are para-linguistic in nature: they involve language but are clearly not directly related to language any level of speech variation.

LANGIDs are 16-bit integer constants and consist of two parts: a primary language identifier and a sub-language identifier. In general, it is sub-language identifiers that are used to distinguish among categories that are not language related. Usually, processes reference the complete LANGID as a whole, however.

Many of the LANGIDs are, in fact, distinguishing locales, even though a separate locale identifier (LCID) mechanism is provided. Consider, for example, the constants 0x0410 "Italian-Italy," and 0x0810 "Italian-Switzerland." As far as we know, Italian as used in Italy is not a different language from Italian as used in Switzerland, nor are we aware of significant differences that would be relevant for linguistic processing. Rather, these two constants are provided to identify different cultural conventions used in the two countries, such as date and number formats, and currency symbols. Such distinctions are made for several major European languages. For example, there are 18 constants for English in different countries, 20 for Spanish, 16 for French. These distinctions are made on the basis of the sub-language identifier component, and in some of these cases, the distinctions may certainly be true sub-language distinctions; for example, different spelling checkers would be needed for UK English and US English. Yet this is clearly not a factor in every case, as in the example of Italian.[3]

[3] One could maintain that the distinctions reflected here correspond to different dialects that might need to be distinguished for speech recognition or speech synthesis, but an examination of all of the identifiers in question suggests that this is not likely the case for every distinction. Furthermore, it is fairly certain that these constants were not defined in anticipation of such use.

That LANGIDs (actually, sub-language identifiers) are really being used to distinguish cultural conventions, which are usually associated with locales, is evident from examining the structure of LCIDs: these are 32-bit constants that are made up of a LANGID and a sort ID. Except in cases in which more than one sort order is defined for a language, LCIDs are simply zero-extended LANGIDs. Thus, locale distinctions are made primarily by the LANGID.

A number of the LANGIDs make distinctions between writing systems used for specific languages.[4] Consider, for example, 0x0450 "Mongolian (Cyrillic)" versus 0x0850 "Mongolian (Mongolian)." Also, in at least one case (0x040A "Spanish-Spain (Traditional Sort)" versus 0x0C0A "Spanish-Spain (Modern Sort)"), two constants are used to distinguish two sort orders, which are an aspect of orthographic conventions.[5]

In one instance, 0x0812 "Korean (Johab)," a LANGID appears to have been created to identify a specific codepage (cp1361). While this LANGID is still documented, Microsoft apparently discontinued implementing it in their operating systems after September 1997.

Where such inconsistencies in definition occur within proprietary systems, it may not be a problem if that system is intended to support only a limited number of languages and if the identifiers are not exposed to end users. For instance, that LANGIDs are used to distinguish locales would not be important if they are only used internally. These identifiers *are* exposed to end users, however; for example, Microsoft Word 2000 allows a user to tag runs of text using any of the LANGID categories. This introduces potential for confusion, particularly when several orthogonal distinctions (codepages, writing systems, sort orders, spelling conventions, other cultural conventions) are overloaded in a single LANGID. This would be especially problematic if the system were extended to cover a large number of languages.

5.2 Systems that identify language-related categories of varying types

Some systems include language identifiers that correspond to different types of language-related categories (languages, dialects, language families, etc.) reflecting inconsistent operational definitions or a lack of such definitions altogether. This occurs, for example, in ISO 639-2.

ISO 639-2 provides codes both for languages and also for groups of languages. Examples of the latter include *cmc* "Chamic languages," and *bnt* "Bantu (Other)." In many of these cases, that a language grouping rather than an individual language is intended is made clear by including either "languages" or "(other)" as part of the name. This does not occur in all cases, however. So, for example, *nah* "Nahuatl" and *que* "Quechua" are presented as though they represent individual languages. There is no question, however, that multiple, distinct languages are being represented in each case. This may not have been evident when these codes were first introduced in bibliographic systems, but today these are undisputed, linguistic facts.

The language groupings in ISO 639-2 are also not consistent in their definition: in most cases, such as *cmc* and *bnt* (mentioned above), the group corresponds to a generic classification of linguistically-related languages—a language family. There are a number of groupings, however, such as *sai* "South American Indian (Other)," which are linguistically heterogeneous, including languages that are not linguistically related. But even *cmc* and *bnt* are not comparable, because *cmc* refers to an entire language family, but *bnt* only refers to the members of the Bantu family that do not have their own code (for instance, it excludes Swahili).

We have acknowledged that the use of identifiers for groups of languages may be appropriate for bibliographic cataloging, which is the context in which the ISO codes originated. But this standard is being used within IT as a general-purpose system of language identification, including identification that is needed for linguistic processing. Again, generic language classifications are not useful for this purpose.

---

[4] Cases of "language" identifiers distinguishing writing systems are also found in the Macintosh operating system.

[5] It is not clear to us why this distinction between sort orders was not made using the sort ID component of LCIDs.

Also, in a system that is primarily intended to identify languages, it may be important for a linguistic process to know when a category that is listed corresponds to a different type of category, such as a dialect or variant. The inconsistency in the types of categories used in ISO 639-2 combined with the lack of definition and the failure to consistently identify the type of each category listed presents significant problems in relation to linguistic processing.

While the problem described with Win32 LANGIDs pertains to a proprietary system used in a single vendor's software, ISO 639-2 is a public standard, used in many IT systems and as the basis for other important IT standards, such as RFC 1766. For this reason, we consider the problem of definition with ISO 639-2 to be of greater concern, and one that also applies to RFC 1766 by inheritance.

The introduction to ISO 639-2 states that the language identifiers it provides are "devised for use in terminology, lexicography, information and documentation (i.e. for libraries, information services and publishers) and linguistics" ([4], p. iv). These represent a fairly divergent set of domains with somewhat differing needs. Most of the codes defined in this standard were originally developed for bibliographic cataloging, and we have already seen that the need for identifiers for this domain does not match the need for identifiers for linguistic processing. At the very least, careful consideration must be given to the issue of operational definitions if we are to continue using this standard as a basis for language identification in technologies such as XML. It is preferable to distinguish language identifiers used for bibliographic subject cataloging from identifiers used for linguistic processing. As mentioned earlier, it may in general be preferable to identify distinct namespaces of language identifiers based on different definitions to be used for different purposes.

6. The problem of scale

Regardless of the exact operational definition used to categorize languages, there is no debate that the number of languages spoken in the world today is very large, on the order of several thousand. The *Ethnologue* catalogs over 6,800 living languages spoken in the world, based on a primary criterion of mutual non-intelligibility. Existing systems of language identifiers list only a fraction of these languages, however. For example, ISO 639-2 lists roughly 400 languages.[6] Thus, the actual number of languages is at least an order of magnitude greater than what existing systems cover.

The need for systems to cover thousands of languages is real, not merely hypothetical. For instance, SIL has been involved in projects in some 1,600 different languages, of which about 1,100 are current, and new projects are begun regularly. Thus, just within SIL, we have an immediate need for over 1,600 identifiers that conform to RFC 1766 for use within XML documents. We are aware of several other agencies that have similar, vastly multilingual needs, such as the Linguistics Data Consortium, the Linguist List, the Endangered Language Fund, UNESCO, various departments of the U.S. and other governments, and others. When we add the work of other institutions, individual linguists and the language communities themselves, the existing needs for language identifiers are considerably greater, and are only continuing to grow. As stated earlier, every language in the world represents a real need for a unique language identifier.

When confronted with needs for thousands of language identifiers, we find that some existing systems do not scale well. There is the obvious problem of devising several thousand new tags. There are other problems with scaling, however, due either to the mechanism that a system uses for tags, or to the procedures for extending the coverage of a system. We will consider each of these in turn.

---

[6] Establishing an exact count of languages is not easy: There are currently a total of 437 codes defined in ISO 639-2/B. On a quick analysis, we found that 56 of these explicitly denote language groupings (e.g. *alg* "Algonquian languages"). There are 381 codes that are presented as denoting individual languages, but a yet-to-be-determined number are in fact groupings (e.g. *que* "Quechua"). At least 16 are for non-modern languages (e.g. *non* "Old Norse") and at least 3 are for artificial languages.

6.1 Mechanism limitations

Some existing systems use mechanisms that cannot allow for thousands of language identifiers. One example of this is Win32's 16-bit integer-based LANGID constants. Because of the way these are implemented in software, there is an upper limit of 512 primary language identifiers that Microsoft can define.[7] This falls well short of existing needs, let alone future needs. It is true that each primary identifier can be combined with 32 sub-language identifiers, but utilizing such combinations in order to cover thousands of languages would involve significant difficulty in managing the assignment of identifiers. Furthermore, it would require abandoning the model assumed by the primary versus sub-language distinction, which would in turn be complicated by current use of sub-language identifiers to distinguish locales and other para-linguistic categories. There is no practical way to make this system expand to cover several thousand languages.

The mechanism used in ISO 639-2, which is to use sequences of three alphabetic characters, has the potential for over 17,500 tags to be defined. The committees that manage this standard, however, place a high value on devising tags that are mnemonic, bearing a resemblance to the name of the language. This can be made to work reasonably well for a few hundred languages, but it is not possible to maintain for thousands of languages: there are simply too many cases of names that are similar. We believe that there is no reason why being mnemonic should be mandatory, and reasons why this may not always be preferable (which we consider in section 9.2). If the committees responsible for ISO 639-2 do not make this a requirement for all tags, then this mechanism can easily scale; otherwise, it will not.

6.2 Procedural limitations

Perhaps the more significant concern with regard to scale has to do with procedures. This is a concern particularly for ISO 639-2. In general, ISO standards involve a relatively lengthy process of review. More significantly, the rules governing additions to ISO 639-2 require that anyone submitting a request for a new language identifier must have at least 50 documents in the language in question. For linguists, for language communities that are just beginning to develop literature in their language, or for agencies wanting to track languages that are not written or only in the initial stages of literacy, this may not be feasible. That does not mean that their need for a language identifier for IT purposes is not valid, however.

The introduction to ISO 639-2 states that the language identifiers it provides are "devised for use in terminology, lexicography, information… and *linguistics*" ([4], p. iv, emphasis added). Linguists need the ability to maintain and exchange linguistic data, including data for languages without literature or extinct languages that may have been documented by researchers but that never developed a literature. ISO 639-2 states that it is intended to serve these users, but its procedures severely limit its ability to scale to cover the diversity of languages that linguists are interested in.

The minimum-document requirement and the careful review of each request for an addition to ISO 639-2 appears to be intended for two purposes: to ensure that the proposed language is indeed of interest to bibliographers, and to ensure that the proposed language is, without dispute, a distinct language worthy of inclusion in the standard. Both purposes raise questions as to whether this standard can meet general processing needs within IT across the full range of application domains mentioned in the text quoted above. In particular, there is an implicit assumption that the standard can arrive at a single, undisputed categorization of languages that are of interest for IT purposes. Given the problem of categorization discussed in section 4, however, that would not seem to be in fact possible. This does not mean that this standard cannot serve important and useful purposes. It is important, however, to understand its limitations and to recognize applications within IT for which it may not be adequate.

We have similar concerns for RFC 1766, which is the basis of language identification in XML. SIL and linguists in a number of agencies are developing systems built on XML for storing, processing, archiving

---

[7] The Macintosh OS also uses 16-bit integer-based language constants, but does not use a two-tier model. Thus, it can accommodate thousands of languages.

and publishing linguistic data. They need to be able to tag elements with identifiers for any of the world's 6,800 languages. As a result, there are immediate needs for tags that conform to RFC 1766 (or its eventual successor).

This standard does not impose the same requirement for 50 documents that is required by the ISO standard. Nevertheless, we are concerned about the ability of the procedures used within RFC 1766 to cope with thousands of requests for new tags and to be able to maintain such a large number of identifiers while avoiding overlapping and redundant identifiers.

7. The problem of documentation

Language identifiers as used in IT are merely labels for identifying and distinguishing speech varieties. As such, they are in principle of very little use unless users are able to determine what each identifier denotes, i.e. what is the location and range of variation in speech that is represented. Yet the systems currently in use for IT either fail to document this at all, or fail to do it adequately.

Both ISO 639(-x) and Win32 LANGIDs, for example, provide no identification beyond a simple name. This creates numerous problems for users. Consider these examples in relation to ISO 639-x:

- Different languages can have similar names, and so can be confused even though ISO 639-2 may provide identifiers for both. For instance, "Chippewa" is an alternate name for "Ojibwa", but these are distinct from "Chipewyan", a language from a different family. A user encountering the code *chp* "Chipewyan" could easily mistake it for "Chippewa", however. There is no way for the user to determine what the intended relationship is between the ISO 639-2 codes *chp* "Chipewyan" and *oji* "Ojibwa" since no identification of what these categories denote is provided.

- In several cases, ISO 639-2 provides a code for one particular language, but if the name is used by several distinct languages, it is not clear which is intended. For instance, the standard tells us that the code *bin* corresponds to a language name of "Bini", but this name is used for different languages, including a Niger-Congo language of Nigeria alternately known as "Edo", and an Australian language alternately known as "Pini". This name has also been used to refer to particular dialects of the Yoruba and Anyin languages of West Africa. There is no additional documentation to tell us which of these (or perhaps some other candidate) is intended. This is only one of many instances of this problem.

- Since ISO 639-2 has codes both for individual languages and also for groups of languages, there is potential confusion for users as to the intended meaning of the codes for groups. For example, the code *ath* denotes "Athapascan languages". Navajo is a particular Athapascan language, yet it has its own code, *nav*. It is not made clear to users whether "Athapascan languages" is intended to include *all* languages from that family, including Navajo, or only those that do not have their own code.

- Because ISO 639-x has categories for individual languages and also has categories that are presented as though they are individual languages but which in fact are groups of related languages, it may be unclear to users what is intended. For example, it is unclear whether *ara* "Arabic" is intended to refer only to Standard Classical Arabic, or whether it is intended to also include all of the various, regional vernacular Arabics, which are distinct languages in their own right.

- For each category in ISO 639 part 1, the two-letter codes, there is supposed to be a corresponding three-letter code in ISO 639-2. But different operational definitions appear to have been used in the two parts of the standard, with the effect that the three-letter codes are more finely grained. So, for example, there is a single two-letter code *st* for "Sesotho". Yet this category is covered in ISO 639-2 by two distinct categories, *nso* "Sotho, Northern" and *sot* "Sotho, Southern". There is currently nothing to indicate to users whether the two-letter code is to be equated with specifically one or the other of the three-letter codes, or whether it is to be understood as the union of the two.

Similar problems arise for Win 32 LANGIDs. For example, it is unclear what 0x0430 "Sutu" is intended to denote. The *Ethnologue* identifies "Sutu" as an alternate name for a Bantu language known as "Ngoni" that

is spoken in Tanzania, Malawi and Mozambique by a combined population of nearly 1,000,000. "Sutu" is also close to "Sotho," used for two other Bantu languages from a different branch of the family: "Northern Sotho," spoken in South Africa and Botswana by some 4,000,000 speakers, and "Southern Sotho," which is spoken in South Africa by over 4,000,000 speakers. Furthermore, the *Ethnologue* indicates that the latter is alternately known as "Suto" or "Suthu." Although the name used by Microsoft, "Sutu," exactly matches one of the names for the language also known as "Ngoni" and does not exactly match any of the names for "Northern" or "Southern Sotho," an examination of the facts suggests that it is most likely intended to refer to the combination of "Northern" and "Southern Sotho," which have a combined population of over 8,000,000. There is no way to be certain of this, however, since Microsoft does not provide any clues to the intended identity.

Likewise, Win32 has several LANGIDs for Arabic, each specific to a particular country. Yet we have seen that LANGIDs are, in general, making locale distinctions. Thus, it is not clear whether these should be used to distinguish among regional vernacular Arabics, or whether they all refer to Standard Classical Arabic but with differences in cultural conventions (e.g. date formats).

When systems fail to document what the categories provided denote, very serious problems will result if the systems are extended to cover a large number of languages. Many languages are known by a variety of names, and cases of potential ambiguity are likely to abound.

RFC 1766 does somewhat better in this regard in that it specifies that requests for new tags should be accompanied by bibliographic references to identify what the proposed tag is to denote. This is an improvement, but is still not entirely satisfactory. For some languages, there may not be any works available that describe the language, and there also may not be any works available in the language that can provide examples of the language.[8]

Even in cases where works exist that do describe the language, they may not be readily available to potential users of the tag, and the likelihood of availability may be very significantly reduced after a number of years have elapsed. (Bear in mind the likely possibility that this system may still be in use fifty years from now and beyond.) Also, after many years have elapsed, the denotee for the tag may have changed: the language community may have shrunk considerably; they may have migrated to a different location, possibly in a different country; or they may have separated with a sizeable portion emigrating to another country on another continent. In these circumstances, it becomes more difficult for users to determine how tags should be used, given only outdated references.

Ideally, what is needed is an online repository of encyclopedic information about the denotation of language identifiers that is maintained and updated on an on-going basis. Of course, this would not be an easy resource to provide.

8. Solving the problems

We have identified problems for language identification in several areas:

- the dynamic nature of languages and of our knowledge of languages;
- the need for operational definitions of language, and the possible need for different categorizations of the world's languages based on different operational definitions devised for different purposes;
- existing systems generally fail to use consistent operational definitions or to clearly indicate distinctions between identifiers for different types of categories;

---

[8] At the time of writing, consideration is being given to having a revised version of RFC 1766 allow bibliographic references to be optional, and to allow requests to include descriptive information to identify a language in addition to or in lieu of bibliographic references. The revised standard is still being drafted, however, so it is not clear at this time exactly what provisions it will allow.

- the number of language identifiers needed within IT are at least an order of magnitude greater than what existing systems already cover, and these systems may not be able to scale; and
- language identifiers must be documented with information that enables users to determine what the identifiers are intended to denote.

We have argued that the operational definition on which a categorization of languages is based must be stated clearly, and have shown that existing systems of identifiers have significant problems in this regard. Given that alternate operational definitions are possible, and that these generally correspond to differing needs, the IT industry must decide among alternatives to deal with this: to maintain the status quo, which is to effectively ignore the problem; to select a particular operational definition to be set apart for use in IT; or to allow for alternate operational definitions in an organized manner, such as allowing for alternate namespaces of identifiers that correspond to alternate definitions that are appropriate for use in IT. In general, it is preferable that a namespace of identifiers be based on a consistent operational definition, and that it should always be clear to a user what type of category (what operational definition) applies to a given identifier.

There are real needs to provide identifiers for each of the world's thousands of languages. In order to deal with the scale of the need, a system must employ a tag mechanism that supports thousands of identifiers. A system must also employ procedures that are able to handle large volumes of requests quickly and efficiently, that are able to accept requests for unwritten languages, and that are able to manage large numbers of identifiers on an on-going basis. In particular, there is an existing need for thousands of additional identifiers to be made available for use in XML.

For the problem of documentation, we have suggested that users must have easy access to encyclopedic information that identifies the speech variety denoted by any given identifier, and that this be available indefinitely into the future. Ideally, this would take the form of an on-line repository of information about the languages denoted by the identifiers listed in a system. Also, given the dynamic nature of language, this information should be maintained and updated on an on-going basis.

9. A specific proposal

A solution to these problems would be considerably advanced by a compilation of language information

- that consistently applies an operational definition of language so that all entities for which an identifier is assigned are of a comparable nature,
- that encompasses all of the languages of the world,
- that clearly documents the speech variety that each identifier denotes,
- that is maintained and updated on an on-going basis, and
- that is freely and readily accessible to the public over the Internet.

In this regard, we propose the *Ethnologue* as a resource that has these properties and that can provide us with significant advances in addressing the problems we have described.

As mentioned earlier, the *Ethnologue* is a catalog of the world's languages that has been compiled by SIL over a period of more than fifty years. Published editions have regularly appeared since 1951. Now in its 14th edition, the *Ethnologue* lists over 6,800 distinct living languages based on a primary criterion of mutual non-intelligibility. Consistency in application of the criteria for identification is maintained by an editor who makes reference to the latest published sources and who solicits the latest information from researchers in the field. This has been a full-time position for approximately thirty years.

The entries in the *Ethnologue* are listed country-by-country, each entry giving basic facts about the use of the language in that country. The most commonly-provided pieces of information include alternate names, names of dialects, number of speakers, geographic location, and linguistic classification. Many languages

are spoken in more than one country, so there are in fact nearly 9,000 entries, but all entries that refer to the same language are linked together by sharing the same language identification code.

The *Ethnologue* allocates a unique three-letter code to each language, and these codes are accessible to any user of the World Wide Web where the full text of the *Ethnologue* is available on the SIL Web site. This is already a heavily-used Web resource; for instance, it serviced 410,000 page requests in May 2000. In order to discover the three-letter *Ethnologue* code for a particular language, one can jump straight to the search page at http://www.sil.org/ethnologue/search/. On typing a language name into the search box, the system will list all the entries in which that name occurs as a main name, alternate name, or dialect name. For instance, if we search for "Hopi" (a language spoken in Arizona by over 5,000 people), we discover that its three-letter code is HOP. One can also search on the name of a country or region and retrieve a list of languages for which the *Ethnologue* entry makes reference to that location.

9.1 The *Ethnologue* in relation to ISO 639-2 and RFC 1766

For someone who needs a unique code today to identify one of the 6,000-plus languages not covered by ISO 639-2, these *Ethnologue* codes are already available. For instance, the XML recommendation ([7], section 2.12) specifies that the value of the language identification attribute must be a language tag as defined by RFC 1766. The RFC mandates that tags not registered with ISO or IANA begin with the primary language tag "x". Thus, the following would be a valid (though private) way of indicating that the content of the `<body>` tag is in the Hopi language:

```
<body xml:lang="x-hop">
```

Even more descriptive would be to add a sub-tag to indicate that SIL is the source of the code. For instance,

```
<body xml:lang="x-sil-hop">
```

On seeing this, a user familiar with this convention could perform a lookup at the SIL Web site to find out what language is denoted by the tag. A CGI script for this purpose is already running. For example, the URL to find out what language is denoted by the *Ethnologue* code HOP is the following:

```
http://www.sil.org/ethnologue/lookup?hop
```

While using private-use codes beginning with "x-" is valid, it is not ultimately satisfactory, particularly for government or major development and relief agencies. Members of the IT community who need to identify languages would prefer to use codes that are widely known and publicly sanctioned, either by inclusion in ISO 639(-x) or by having a tag registered with the Internet Assigned Numbers Authority (IANA), in accordance with RFC 1766. The set of languages of interest to members of the IT community include all those listed in the *Ethnologue*, and so, to this end, we would like to see the entire inventory of languages listed in the *Ethnologue* added either to ISO 639-2 or to the IANA registry.

In either case, there are potential problems with the registration process scaling to cope with the volume of tags in question. A simpler procedure that could be considered would be a tag "i-sil" to be registered under RFC 1766 that would indicate that the following sub-tag is a three-letter code maintained by SIL in the *Ethnologue*.

A tag of "i-sil" easily deals with problems of scale: it allows for a simple procedure to be used to introduce the thousands of new languages needed, and also provides the actual tags without any need for each one to be separately devised.

Such a tag can also be helpful in solving the problem of alternate categorizations: used as suggested above, it would, in effect, identify a distinct registration authority for language identifiers that provides its own namespace of tags based on a specific operational definition. This option need not be limited to the *Ethnologue* but could be made open to other agencies that met some established criteria (including a statement about operational definitions) and that provided significant benefit for users. Current IT implementations assume that a single namespace for language identifiers is to be preferred, and there certainly are valid arguments in favor of that, but these are perhaps offset by the benefits in terms of addressing the problems of categorization and scale. Concerns about the use of alternate namespaces can

be reduced by a requirement that an ISO code would take precedence over a code from an alternate namespace if an ISO code is defined and if the denotations of the ISO and alternate codes are identical. A requirement of this nature is already included in RFC 1766 in relation to IANA-registered tags. Furthermore, we assume a requirement that an agency maintaining an alternate namespace would be required to define a mapping between tags in that namespace and codes in ISO 639-x.

A mechanism for managing alternate namespaces could be further refined by introducing into RFC 1766 (or a successor) a new primary tag such as "n-". This would designate that the following sub-tag identifies a registered namespace authority for language identifiers. This would imply that the tags that follow are documented and maintained by that namespace authority. The primary tag "i-" becomes a privileged namespace identifier denoting IANA as the authority for that namespace. Each other namespace authority would register an identifying tag with IANA. One benefit of this is that tags beginning with "i-" would continue to designate only language categories, and not also namespaces of language tags.

9.2 Possible objections

In recent years the idea of using the *Ethnologue* codes as a system for language identification has been suggested by a number of people on various email discussion groups. This has always elicited responses both favorable and unfavorable. The following are the major objections that have been raised and the way we would answer them.

The codes are not mnemonic enough.

The system of three-letter codes allows for a total number of 17,576 possible codes. Thus, the current inventory of about 6,800 codes does not even use half the available codes. Enough of the codes are in use, however, that a large number do not bear a mnemonic resemblance to the name of the language they denote. Because there are very many cases in which different language have similar names, maintaining a requirement of mnemonic resemblance for three-letter codes quickly becomes impossible.

Some have complained about the fact that many *Ethnologue* codes are not mnemonic and have proposed that, as allowed by RFC 1766, we could use up to eight letters in a code in order to make them more mnemonic. We have found the three-letter system adequate, however, and see at least two disadvantages of adopting longer codes. First, many language names are significantly longer than eight characters, and there would be numerous ways of abbreviating them to eight characters. Remembering the precise spelling of an eight-character abbreviation could actually be a greater burden to memory than remembering the more arbitrary three-letter code. Secondly, languages are often known by several names, and mnemonic names may conform to only one. Thus, a decision between alternate names must be made in devising the tags, and the resulting tags are not mnemonic for all users. More importantly, alternate names often carry political overtones. Whereas an arbitrary three-letter code can be devoid of such overtones, a mnemonic eight-letter code would end up promoting one faction's point of view over the objections of another. Similarly, as the name considered most appropriate may change over time, a mnemonic eight-letter code could end up endorsing an inappropriate name from the past.

The system of codes is too detailed.

Many are overwhelmed by the prospect of over 6,800 language codes: "Isn't there a way we could get by with fewer codes?" If the basic motivation for language tagging includes the need to associate linguistic data with the correct behaviors for language-specific processing, then this is approximately the number of codes we will ultimately need since this is about how many mutually non-intelligible varieties of speech are used in the world today. For instance, for each language denoted by one of the 6,800 codes, if someone wanted to spell-check text in that language, they would need a spelling dictionary specific to that language. This level of granularity generally corresponds to the level at which distinctions relevant for linguistic processing occur. One might argue that most of these language groups are so small as not to be significant from a global perspective, but we reiterate that every one of these languages is significant to some portion of the world's citizens who will need a way to identify it when they encode it in a data stream.

The basis for defining language is not right.

We have already discussed the issue of operational definitions of language at length. The definition we have chosen, based primarily on the criterion of mutual non-intelligibility, is one that identifies speech varieties of broad interest in linguistic research and that is widely accepted among linguists, that identifies speech varieties of interest to development agencies for purposes of literacy and language development, that generally corresponds to the level at which distinctions that are relevant for linguistic processes occur, and that has served our institute well. The *Ethnologue* has not made clear in the past what operational definition it has assumed, but we have recognized the need to document this better and will make it more clear in the future.

For a general objection of this nature to be convincing, one would really need to present an alternate operational definition and a listing of the world's languages as categorized on that basis. We are not aware of another complete set of language identifiers that is publicly available over the World Wide Web. Though it would be possible to develop a different operational definition for language, it would be such a huge task to apply it consistently across all the speech varieties in use in the world today that it is not likely to happen soon.[9] Furthermore, should an alternate set based on a different operational definition become available, we have suggested that, in principle, it may be preferable within IT to allow alternate namespaces of identifiers based on alternate operational definitions.

On various occasions, objections have been raised regarding the categorization for a particular group of related languages; for example, "There are dozens of listings for Quechua." Statements like this often imply that there must be errors. These objections often reflect a lack of familiarity with the languages in question and with the fact that several mutually non-intelligible speech varieties are, in fact, represented. In some cases, confusion results from there being one dominant language or one written form that each of these languages is associated with. In these cases, the objections reflect different assumptions about operational definitions. It is possible that the information on which the *Ethnologue* is based is incomplete or not entirely accurate, but there are avenues available for submitting corrections. For such specific objections to stand, one would really need to present documented evidence that uses the same operational definition based on mutual non-intelligibility but that points to a different categorization for these languages.

The Ethnologue *contains inconsistencies and is not mature enough to serve as the basis for a standard.*

On various occasions, people have obtained an electronic copy of *Ethnologue* data and have processed that data, and then have objected that there are numerous duplications and inconsistencies. Such claims arise mostly from not understanding the organization of the data they have obtained. (In this regard, SIL has been at fault for not adequately documenting what the organization of the data is.) In the past, the *Ethnologue* data was maintained in a flat, text-file database with one record for each language in each country in which it is spoken. Furthermore, in many cases, different records for a given language would give a different language name based on preferences in that country. Within the *Ethnologue*, it is not a particular name that distinguishes a language from every other, but rather the three-letter codes. The perception of duplication and inconsistency is an unfortunate by-product of the view into the data.

The use of a flat, text-file database did indeed result in some inconsistencies in the way of spellings, population counts, bibliographic references, etc. This is quite a different problem from that described above. The *Ethnologue* database has been undergoing a major overhaul, however, and will be maintained in the future as a relational database. This will address both problems mentioned here: it provides validation mechanisms and ensures consistent spellings, population counts, etc. It also makes it possible to

---

[9] The recently published *Linguasphere Register* (http://www.linguasphere.org) is the most likely candidate in this regard. It is not freely available over the Internet, however, and we have not yet been able to examine it to see precisely what operational definitions it employs and whether these are suitable for use as a basis for language identification in IT. That research effort appears to have been pursued over a comparable length of time as the *Ethnologue* (almost 40 years for the *Linguasphere*, and over 50 years for the *Ethnologue*). Comprehensive research of this nature simply takes a very long time, and we are not currently aware of any other potential candidates.

present many alternate views of the data, including views that do not suffer from perceptions of duplication or inconsistency in the categorization.

The Ethnologue *lists distinct languages where one common form is what is of interest for IT processing.*

The typical observation being made here is that several languages may share a common written form, and it is only that form that matters for certain IT processes. For example, the *Ethnologue* lists several speech varieties in Italy that are closely related to Italian, but some maintain that there is only one language of interest.

This merely supports our observation that different categorizations of languages are possible, based on different definitions and for different purposes. For one group of users, if they are only concerned about the linguistic variety represented, say, in major Italian newspapers, then an identifier should be available for their use. But another group of users may be interested in the various, mutually non-intelligible speech varieties reflected in the *Ethnologue*, and identifiers should also be available for them.

It should also be pointed out that this type of objection is typically associated with languages that have long histories of literacy combined with expansion and diversification of the language community. When considered in relation to the totality of thousands of languages spoken in the world, these cases are more likely to be exceptions than norms.

The Ethnologue *lists one category where several distinctions are needed.*

It is true that many processes, such as spell-checking, require a finer level of granularity than the categorization provided in the *Ethnologue*. As pointed out in footnote 1, many processes actually depend upon para-linguistic distinctions, such as writing system or spelling conventions, which are different from language per se. The variety of such notions that may be needed for IT processes, and which process depends on what notion, is a topic that we believe needs further consideration within IT. It appears clear, however, that enumerations of categories for such para-linguistic notions must build off a categorization of languages per se; a list of *languages* is logically prior to lists of writing systems or spelling conventions.

This is not an argument against the categorization of languages per se reflected in the *Ethnologue*. Rather, it simply points to the fact that the entire set of issues involving language-related distinctions that need to be maintained for use within IT involves more than just language identification.

The Ethnologue *is oriented toward living languages and does not adequately cover ancient languages.*

It is true that the *Ethnologue* does not meet the needs of users with interests in ancient languages, since ancient languages fall outside its scope. The correct conclusion to be drawn from this is that, if the *Ethnologue* were to be used as the basis of a comprehensive list of identifiers for living languages, then there would be a separate and additional need to arrive at a list of identifiers for ancient languages. (Following the mechanism proposed in this paper, that could constitute its own namespace that would be maintained by its own naming authority.) This does not constitute an argument that the *Ethnologue* is inadequate for coverage of living languages, however.

The Ethnologue *lacks a hierarchical organization for languages.*

The *Ethnologue* does, in fact, give a hierarchical, linguistic classification for every language. What it does not do is reflect such hierarchical relationships within the system of three-letter identifiers that it uses.

Although it may sound like a desirable goal, there are several reasons why attempting to reflect hierarchical relationships within a system of identifiers might not be ideal, at least at this time:

First, just as there is not one agreed-upon categorization of languages, there is also not complete consensus among experts regarding the genetic and historical relationships among languages. There are simply too many unanswered questions with scanty evidence to work from for there to be complete agreement. Yet encoding such a hierarchy in a system of identifiers would require adopting one set of conclusions.

Secondly, historical, genetic relationships represent only one basis for forming a taxonomy of languages. Hierarchies could just as readily be based on other orthogonal issues, such as linguistic properties and similarities, written traditions, etc. There may, in fact, be several such hierarchies that could be of potential interest.

Finally, it is not clear how such hierarchical relationships would be applicable for processing purposes within IT. These relationships may be of interest for bibliographic purposes, for instance, but it is not clear that they have a use in linguistic processing (as defined in section 2). IT systems should not be burdened with such information if it does not serve any useful purpose within those systems. At the least, further research in this regard would be necessary before any comprehensive hierarchical system of identifiers could be adopted.

The codes do not match the corresponding codes in ISO 639-2.

There are two aspects to this issue. First, there are cases in which different codes are used to identify identical languages. For example, for the language Afrikaans, ISO 639-2 uses the code *afr* whereas the *Ethnologue* uses *afk*. Such differences are largely due to having two distinct histories of development. These are not a serious concern since the ISO codes already exist, and either ISO 639-2 or RFC 1766 would give precedence to them. (We are not suggesting that *Ethnologue* codes replace existing ISO codes; only that they supplement them.) Thus, conflicting *Ethnologue* codes for these languages would not present problems for users. We are currently examining exactly what would be needed to provide adequate support to existing users of *Ethnologue* codes. At a minimum, we are prepared to establish and maintain a mapping between these systems and make it publicly accessible from the *Ethnologue* Web site. Work on this has already begun.

The second aspect of this issue has to do with cases in which not merely codes mismatch, but the very categories themselves differ. A clear example of this is the ISO category *que* "Quechua". In the *Ethnologue*, this corresponds to 47 distinct languages (assuming the broadest possible denotation for the ISO term, which can only be assumed in the absence of any identifying information). In cases such as this, we do not see any fault on the part of the *Ethnologue*: it lists 47 languages that correspond to ISO's *que* because these are, in fact, distinct languages according to our consistent operational definition based primarily on mutual non-intelligibility. In contrast, ISO has not used a consistent operational definition. The ISO category *que* is not, in fact, a language but a language family or collection, which is not what is needed as the basis of identification for linguistic processing.

We do not suggest that the ISO codes that identify groups of languages ought to be deprecated. They exist because they have served a need (primarily for bibliographic cataloging), and they have a user base that is not about to abandon them. Rather, what we are suggesting is that the existing ISO codes must be supplemented with additional codes to provide complete coverage for all of the world's languages, thereby meeting needs for identification in relation to linguistic processing.[10]

Again, to assist users in understanding the relationships between ISO and *Ethnologue* codes, we intend to maintain a mapping, as mentioned above, and to make this information readily available over the Web.

If we are looking outside of ISO 639(-x) for sources of codes, the Ethnologue *is not the only source.*

We acknowledge that this is true, and in principle there is no reason why another source could not be used, provided it has a clear operational definition of language that meets the needs of linguistic processing. There is no other source that provides all of the benefits that the *Ethnologue* provides, as listed above, however. Furthermore, other sources do not generally provide the complete coverage that the *Ethnologue* does. Hence, using other sources involves compiling information from multiple sources, comparing the categories listed to ensure that they are of the same type and determining when they have the same or

[10] While we would not argue that generic classifications in ISO 639-2 should be deprecated altogether, we have argued that their use for linguistic processing is inappropriate, and that consideration should be given within IT to the need to distinguish among the ISO 639-2 codes and to determine which are appropriate for use in which applications.

different denotations. In contrast, if the *Ethnologue* is used as a source, this work is not necessary, and codes are already available for use.

The codes are not stable enough to serve as a standard.

There are two distinct but related issues here: the stability of the codes, and the stability of the categorization (the inventory of languages). We will consider each.

A new edition of the *Ethnologue* is published every four years. With each edition there are changes to the overall inventory of languages, and thus the set of three-letter codes changes. For instance, over the past four editions (from 1988 to 2000), the total number of living languages reported in the *Ethnologue* has increased from 6,140 to 6,527 to 6,703 to 6,809. To some extent this reflects the discovery of previously unknown languages, but mostly it reflects the fact that, as we find out more information about the linguistic situation in a region, we often conclude that what were previously considered to be two dialects of the same language are in fact distinct enough (typically on the basis of mutual non-intelligibility) to be considered languages in their own right.

The fact that the inventory of languages and codes changes over time is probably what is being construed as instability. With regard to the codes, we believe that the approach we take to managing the codes does, in fact, make them adequately stable. Our basic rule is this: codes can be retired, but never reused. That is, in the occasional instances when a language is removed from the inventory (for instance, because it is deemed on the basis of improved evidence to be a dialect of another language), its three-letter code is retired from service, never to be reassigned later. Thus a given code would never denote two completely distinct speech varieties over time. The only way the denotation of a code can change is for it to widen in scope when another language is merged with it, or to narrow in scope when the language is split. A new implementation of the *Ethnologue* database keeps track of changes to the status (unused, active, or retired) and coverage (splits and mergers) of codes; a future version of the *Ethnologue* Web site will be able to display the change history for any requested code. In this way, it will always be possible for a user to determine what a code currently denotes, how its denotation has changed over time, and what is the correct code to use to identify a particular language.

As for changes in the categorization itself, it turns out that changes of this nature are likely to be made as a result of improved information only in relation to languages that are not generally of interest to commercial developers of IT systems and their users. As a result, the majority of users are unlikely ever to be affected by any such revisions to the *Ethnologue.*

In addition, as already indicated, such changes in our knowledge of languages are not unlike the kinds of changes that any durable system of language identifiers must accommodate due to the dynamic nature of human languages themselves. In terms of international standards for IT, these kinds of changes are not categorically different from changes that must be allowed for in listings of names of countries and geo-political units, such as ISO 3166 [5].

Furthermore, it should be pointed out that such concerns for instability apply equally to other systems, including ISO 639-2. For example, ISO 639-2 currently has no code for any variety of the Yi language; the closest match is *sit* "Sino-Tibetan (Other)". If a code were added to that standard for Yi (or some variety thereof), then the denotation for the existing code *sit* would change: it would no longer include Yi. These changes are identical to the kinds of changes that occur for *Ethnologue* codes as knowledge of languages is refined. Indeed, in this respect, the *Ethnologue* is superior to ISO 639-2 in that it explicitly documents what each of its codes denotes, and so it is possible to determine when the denotation of a code has changed. Let us examine this further:

Consider, the impact on data by adding a new code to ISO 639-2 for a language like Yi. Currently, if existing Yi data is tagged with an ISO code, *sit* "Sino-Tibetan (Other)" would be the code of choice. Now, suppose after some time a new tag is added for Yi. As pointed out above, the range of languages covered by the code *sit* has suddenly changed since it no longer includes Yi. As a result, the existing data in now *incorrectly* tagged. Furthermore, there is no way in principle to determine the degree of error. Moreover, a user would not even be aware that the data is incorrectly tagged, since all they know about it is that it is *some* Sino-

Tibetan language. If anything, they are more likely to discover that there is a code for Yi, and to be led to the wrong conclusion that the data is in some language other than Yi.

In contrast, the worst situation that would arise from a change in the *Ethnologue* is that data is sub-optimally tagged. For example, suppose some data were tagged to indicate it was in the Akha language using a code based on the *Ethnologue*, such as *i-sil-aka*. Suppose further that it was subsequently decided that this, in fact, corresponds to two distinct languages, with a corresponding change in codes. In this situation, when a user goes to the *Ethnologue* Web site to find the meaning of the code *aka*, they will learn that a new classification based on improved information has been made. Given the date on the document, they will realize that the tag on that data predated the change in classification. The old tagging information still gives them relatively precise indication of the language of the data, however, and they have some indication as to the extent to which the language category denoted by the tag that was used differs from the more current knowledge regarding the language categories. Most importantly, they are able to determine that the tag that was used does not reflect current best practice.

In terms of data maintenance, a limited system of identifiers using generic categories of languages with occasional and careful additions is no better than a comprehensive system of identifiers that undergoes occasional refinement. Indeed, we find that there reasons why the former might be considered much worse.

10. Conclusions

Our motivation for writing this paper has arisen out of an existing need for language identifiers that cover thousand of languages beyond those currently covered by systems such as ISO 639-2. In particular, there is an immediate need for thousands of language identifiers to be used in XML by SIL, various agencies, individual linguists, and speakers of "lesser known" languages around the world. This need must somehow be met.

In considering this problem, we identified a number of problems for language identification in IT in general, and we have described each of these here. Furthermore, we have indicated that these problems are not currently being considered in existing systems of language identification. To our knowledge, these problems have not been identified within the IT industry before now. As a result of the fact that existing systems do not address these problems, we believe that these systems are prone to encounter numerous problems in the future.

We have argued that these general problems, and some specific problems with particular existing systems such as ISO 639-2, must be addressed within the IT industry. Key among these is current lack of any operational definition for language, and the possible need to allow for alternate categorizations based on different operational definitions suited to different purposes. In this regard, we have suggested the possibility of allowing for alternate namespaces of language identifiers that correspond to specific operational definitions.

We have listed several benefits the *Ethnologue* has to offer in providing solutions to these problems. One specific possibility, a tag of "i-sil" registered in accordance with RFC 1766, would immediately address the problem of scale and would provide a set of language identifiers that is:

- based on a consistent operational definition,
- publicly accessible over the internet,
- fully documented as to the denotation of each identifier, and
- maintained on an on-going basis.

Various objections have been raised to using the *Ethnologue* as a basis for language identifiers, but we have presented what we believe are adequate responses to these objections. In addition, we have begun to take steps to provide information that can make the *Ethnologue* of even greater usefulness to users than it

already is, including mappings between *Ethnologue* and ISO codes, and revision histories that track changes in what speech variety a code is intended to denote (e.g. splits or mergers) based on changes in our knowledge of languages. The main concern with using the *Ethnologue* would be to ensure that users understand what it does and does not provide.

Finally, we have suggested that important problems with language identification can be solved by adopting a system of alternate namespaces based on alternate operational definitions. *Ethnologue* codes could constitute one such namespace, with the tag "i-sil" (or "n-sil") representing an instance of an identifier for such a namespace. It would not be possible to arrive at a single standard categorization of all languages that meets the needs of all users. A mechanism for providing alternate namespaces of identifiers based on different definitions would provide a way to move beyond an insoluble problem to meet the diverse needs of users.

11. References

[1] Alvestrand, H., ed. 1995. *RFC 1766: Tags for the identification of languages.* Available online at http://www.ietf.org/rfc/rfc1766.txt?number=1766. (At the time of writing, a revised version is in preparation, and the discussion here anticipates provisions that are expected to be incorporated in the new version.)

[2] Grimes, Barbara, ed. 2000. *Ethnologue: languages of the world. 14th edition.* Dallas: SIL International. The 13th edition is available online at http://www.sil.org/ethnologue/; an online version of the 14th edition is in preparation.

[3] International Organization for Standardization. 1998. *ISO 639:1998(E/F), Code for the representation of names of languages.* Geneva: International Organization for Standardization.

[4] International Organization for Standardization. 1998. *ISO 639-2:1998(E/F), Codes for the representation of names of languages—part 2: alpha-3 code.* Geneva: International Organization for Standardization. Available online at http://lcweb.loc.gov/standards/iso639-2/langhome.html.

[5] International Organization for Standardization. 1997. *ISO 3166-1:1997, Codes for the representation of names of countries and their subdivisions—part 1: country codes.* Geneva: International Organization for Standardization. The current country codes are available online at the site of the maintenance agency: http://www.din.de/gremien/nas/nabd/iso3166ma/.

[6] Microsoft Corporation. 2000. *Microsoft Developer's Network Library.* CD-ROM. Available online at http://msdn.microsoft.com/library/default.asp. The documentation that relates to LANGIDs is found under the heading "Platform SDK, Base Services, International Features, National Language Support, National Language Support Reference, National Language Support Constants, Language Identifiers."

[7] World Wide Web Consortium. 1998. *Extensible markup language (XML) 1.0.* Available online at http://www.w3.org/TR/1998/REC-xml-19980210.