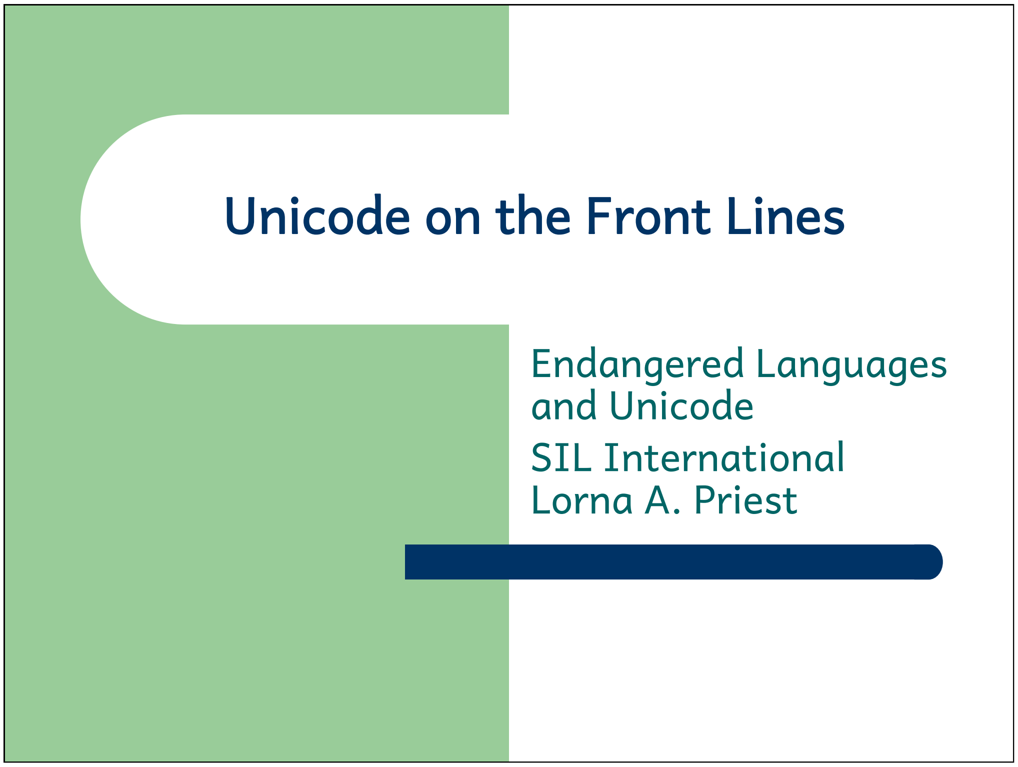


Unicode on the Front Lines

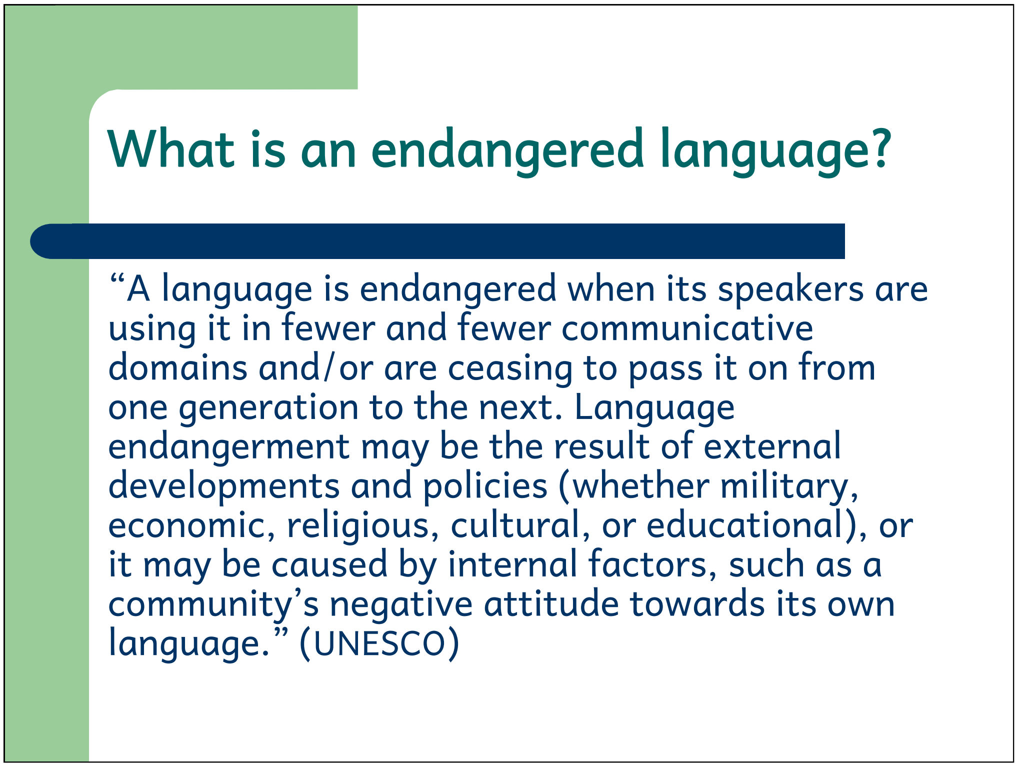
Endangered Languages and Unicode

SIL International
Lorna A. Priest

Introduction

For 70+ years SIL International has been working to study, develop and document the world's lesser-known languages. Most of these languages were previously unwritten. Many of them could also be considered endangered. The loss of many of the world's languages has attracted a lot of attention, particularly in the linguistic world.

SIL's work typically includes academic research, translation and literacy work. As an organization with 2,000+ linguists, coming from over 60 countries, and working in over 1,300 languages in over 90 countries, SIL International had a strong incentive to switch to Unicode encoding. This paper will briefly look at what an endangered language is, how SIL International has been involved in these languages and in particular how our work with Unicode is helping endangered languages.



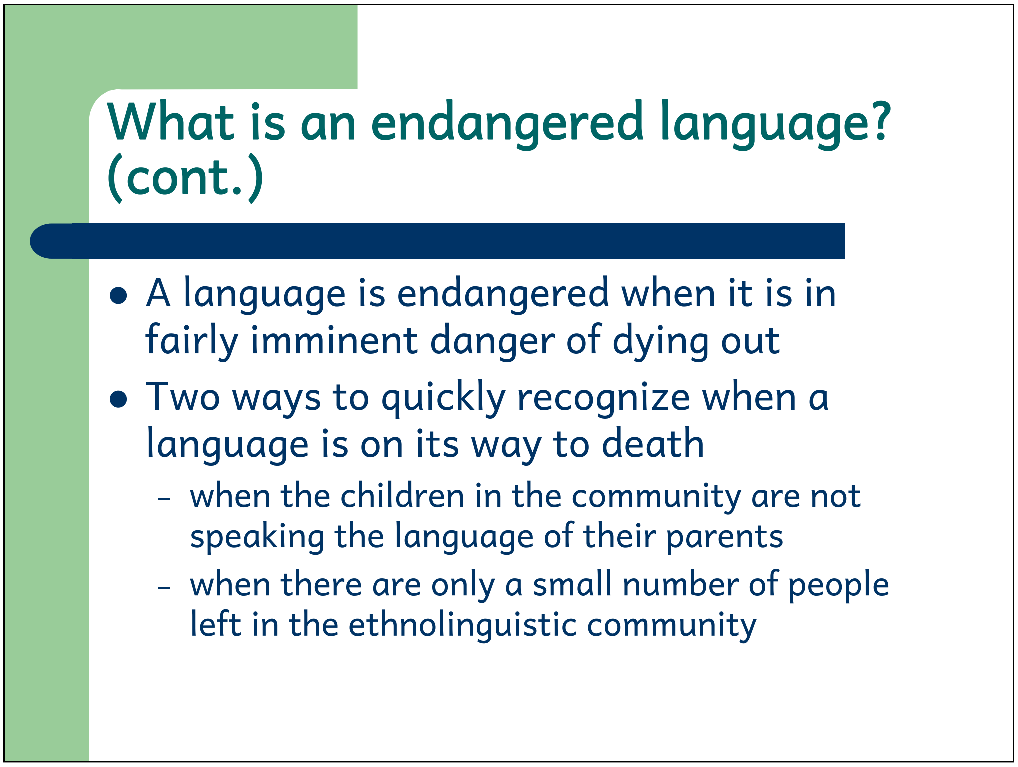
What is an endangered language?

"A language is endangered when its speakers are using it in fewer and fewer communicative domains and/or are ceasing to pass it on from one generation to the next. Language endangerment may be the result of external developments and policies (whether military, economic, religious, cultural, or educational), or it may be caused by internal factors, such as a community's negative attitude towards its own language." (UNESCO)

What is an endangered language?

First, let's look at various definitions of what an endangered language is. UNESCO (page 11) has defined an endangered language as:

> "A language is endangered when its speakers are using it in fewer and fewer communicative domains and/or are ceasing to pass it on from one generation to the next. Language endangerment may be the result of external developments and policies (whether military, economic, religious, cultural, or educational), or it may be caused by internal factors, such as a community's negative attitude towards its own language."

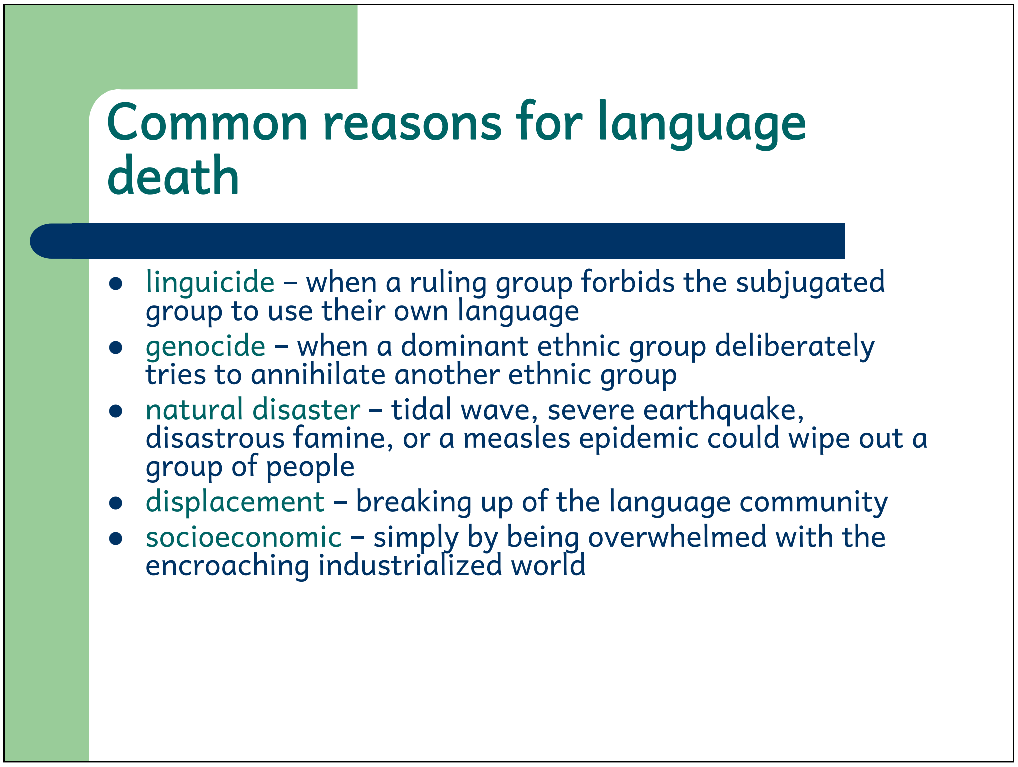


What is an endangered language? (cont.)

- A language is endangered when it is in fairly imminent danger of dying out
- Two ways to quickly recognize when a language is on its way to death
  - when the children in the community are not speaking the language of their parents
  - when there are only a small number of people left in the ethnolinguistic community

What is an endangered language? (cont.)

One of SIL's top linguists, who has chaired the Committee for Endangered Languages and their Preservation of the linguistic Society of America, Michael Cahill, has said, "A language is endangered when it is in fairly imminent danger of dying out". He gives two ways to quickly recognize when a language is on its way to death: when the children in the community are not speaking the language of their parents and when there are only a small number of people left in the ethnolinguistic community.



Common reasons for language death

- linguicide – when a ruling group forbids the subjugated group to use their own language
- genocide – when a dominant ethnic group deliberately tries to annihilate another ethnic group
- natural disaster – tidal wave, severe earthquake, disastrous famine, or a measles epidemic could wipe out a group of people
- displacement – breaking up of the language community
- socioeconomic – simply by being overwhelmed with the encroaching industrialized world

Common reasons for language death

The following are some common reasons for language death:

linguicide – when a ruling group forbids the subjugated group to use their own language

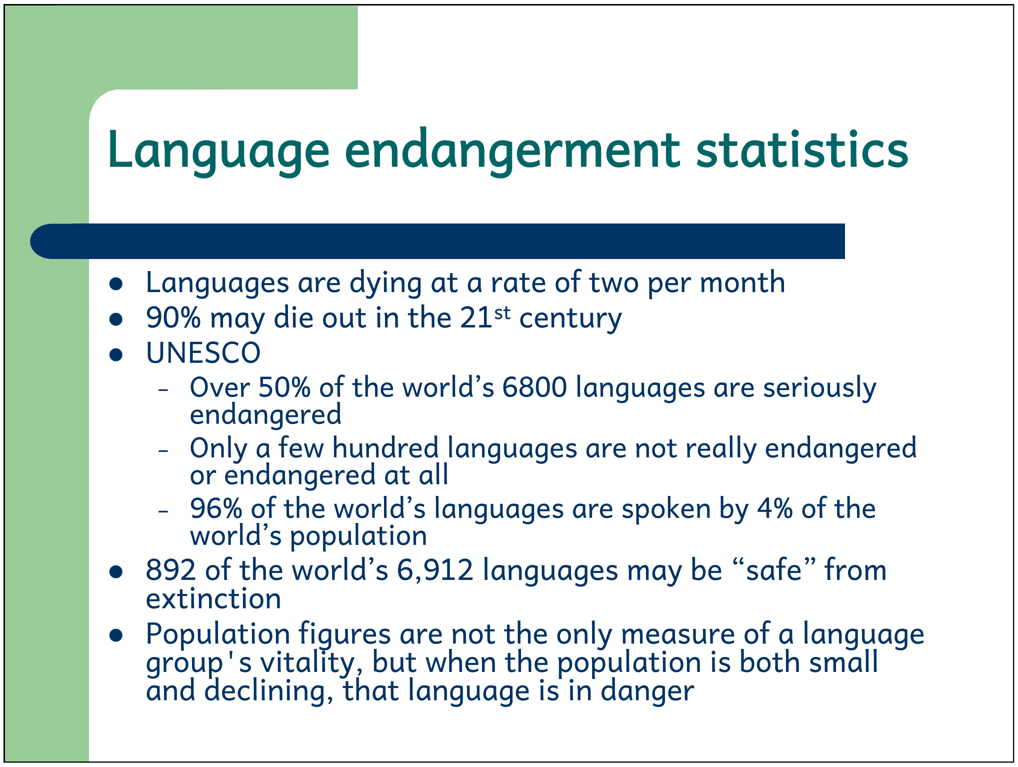
genocide – when a dominant ethnic group deliberately tries to annihilate another ethnic group

natural disaster – tidal wave, severe earthquake, disastrous famine, or a measles epidemic could wipe out a group of people

displacement – breaking up of the language community

socioeconomic – simply by being overwhelmed with the encroaching industrialized world

The main reasons for language death today seem to be as much economic as anything. A parent sees the money and jobs available for people who can speak Language X, which isn't their own language. So they don't teach the kids their own language.



Language endangerment statistics

- Languages are dying at a rate of two per month
- 90% may die out in the 21st century
- UNESCO
  - Over 50% of the world’s 6800 languages are seriously endangered
  - Only a few hundred languages are not really endangered or endangered at all
  - 96% of the world’s languages are spoken by 4% of the world’s population
- 892 of the world’s 6,912 languages may be “safe” from extinction
- Population figures are not the only measure of a language group's vitality, but when the population is both small and declining, that language is in danger

Language endangerment statistics

We all like numbers, and there are many statistics we could quote from. Let us look at a few.

Conservative estimates are that the world’s languages are currently dying at the rate of at least two languages each month. Less-conservative estimates forecast that as many as 90% may die out in the 21st century (Headland, p. 5).

UNESCO (p. 11) statistics say:

-Over 50% of the world’s 6800 languages are seriously endangered

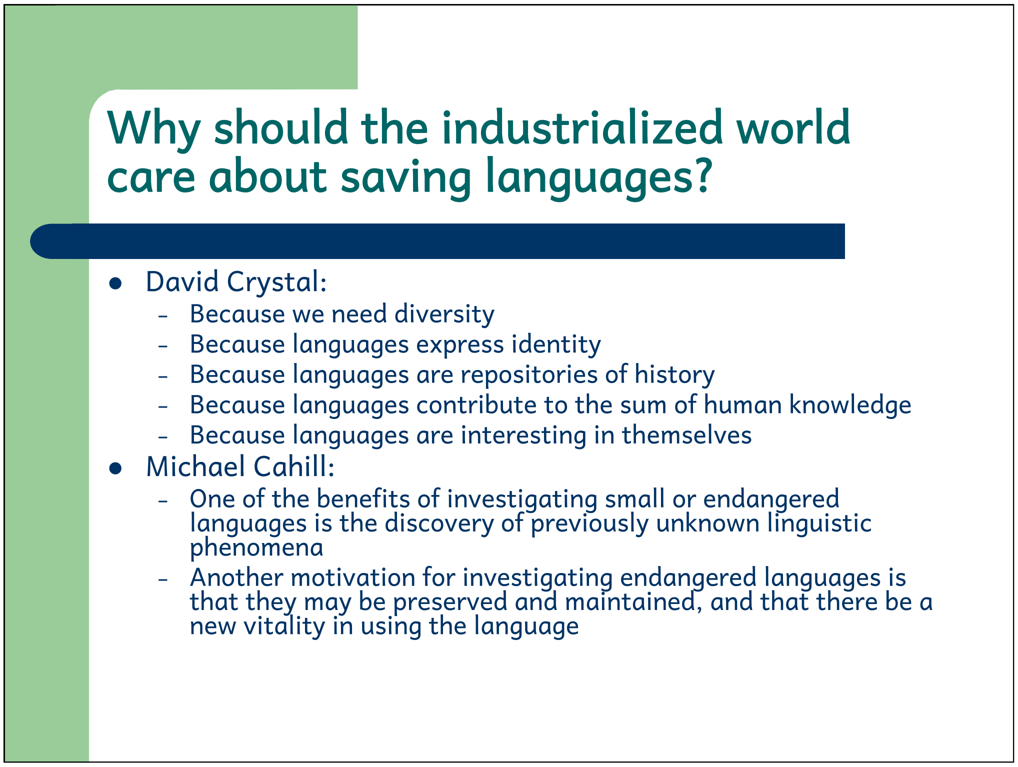
-Only a few hundred languages are not really endangered or endangered at all

-96% of the world’s languages are spoken by 4% of the world’s population

Michael Krauss (p. 7) has said that a language could perhaps be considered “safe” if it has 100,000 or more speakers. Using that figure, and going to the Ethnologue, we see that there are 892 languages with at least 100,000 speakers and which could be considered “safe” from extinction.

Michael Cahill reminds us that “Population figures are not the only measure of a language group's vitality, but when the population is both small and declining, that language is in danger.”

Whatever statistics we use, we see that a large proportion of languages today are not “safe.”



Why should the industrialized world care about saving languages?

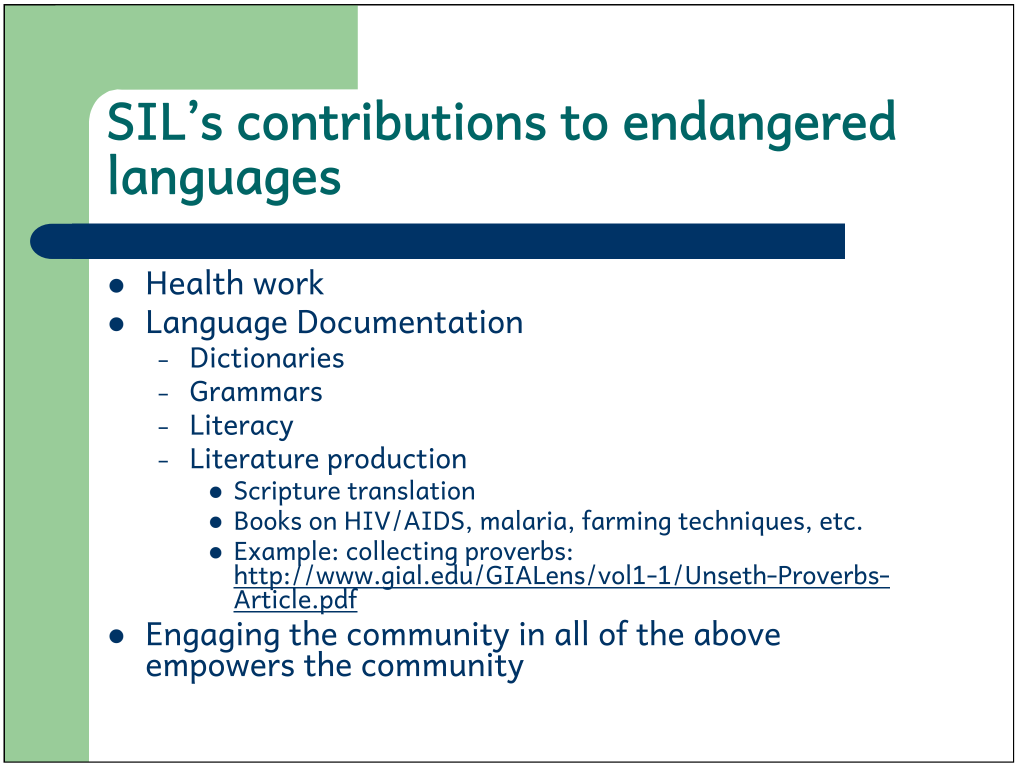
- David Crystal:
  - Because we need diversity
  - Because languages express identity
  - Because languages are repositories of history
  - Because languages contribute to the sum of human knowledge
  - Because languages are interesting in themselves
- Michael Cahill:
  - One of the benefits of investigating small or endangered languages is the discovery of previously unknown linguistic phenomena
  - Another motivation for investigating endangered languages is that they may be preserved and maintained, and that there be a new vitality in using the language

Why should the industrialized world care about saving languages?

David Crystal has written on language death and gives the following reasons why we should care (p. 27-66):

> Because we need diversity – each language gives us a slightly different model of the universe. Today, we have 6,912 models of the universe (Ethnologue says there are 6,912 languages). "If diversity is a prerequisite for successful humanity, then the preservation of linguistic diversity is essential, for language lies at the heart of what it means to be human"
>
> Because languages express identity – language helps every member of the community to experience identity as part of a whole
>
> Because languages are repositories of history – this can be both oral or written. "...once a language is lost, the links with our past are gone"
>
> Because languages contribute to the sum of human knowledge – we can learn a great deal from studying languages
>
> Because languages are interesting in themselves – each language may demonstrate a feature not found in other languages

SIL's Michael Cahill also says "one of the benefits of investigating small or endangered languages is the discovery of previously unknown linguistic phenomena. However, another motivation for investigating endangered languages is that they may be preserved and maintained, and that there be a new vitality in using the language."



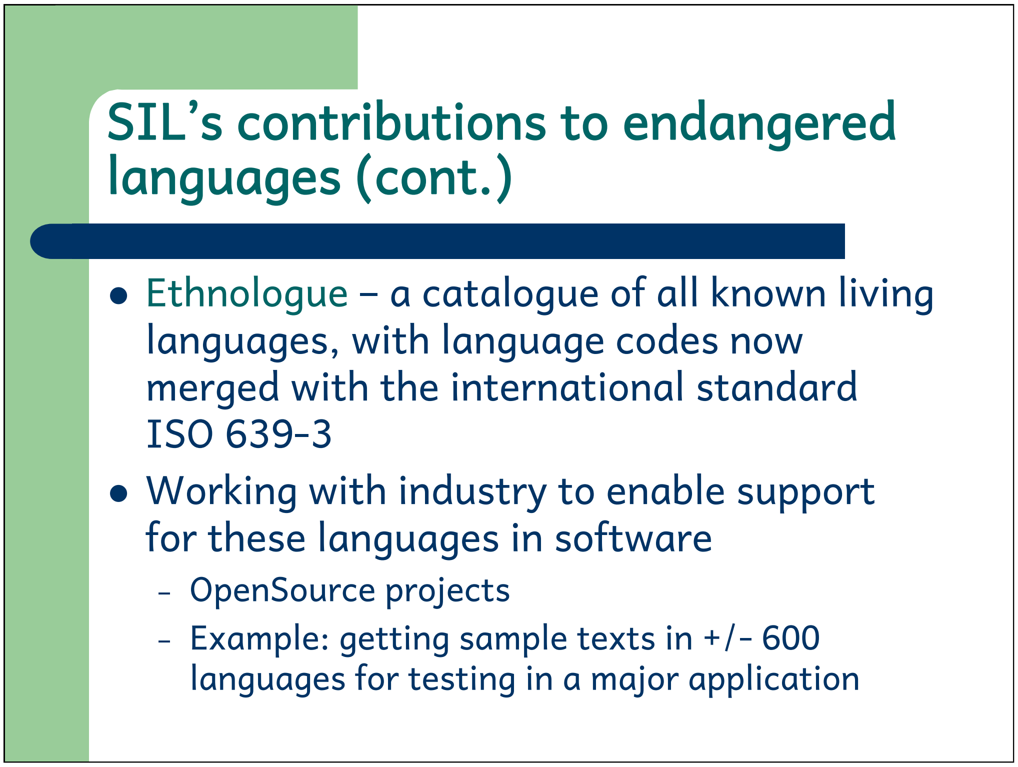
SIL’s contributions to endangered languages

- Health work
- Language Documentation
  - Dictionaries
  - Grammars
  - Literacy
  - Literature production
    - Scripture translation
    - Books on HIV/AIDS, malaria, farming techniques, etc.
    - Example: collecting proverbs: http://www.gial.edu/GIALens/vol1-1/Unseth-Proverbs-Article.pdf
- Engaging the community in all of the above empowers the community

SIL’s contributions to endangered languages

SIL has helped reverse situations where a people group was in danger of dying out. Health work is a major factor in increase of the population because of lowered infant mortality, measles vaccines and tuberculosis treatment.

SIL has been involved in a wide range of language documentation including dictionaries, grammar and literature production. Literacy has been the beginning of a turnaround in the negative perception of themselves and their language. When people can read they feel like their language is worth something. They have a new dignity. It also helps them not be cheated while trading with outsiders if they can read a contract or bill-of-sale instead of relying on someone else to read for them. We have found that self-esteem starts to grow when a people start reading about a God who cares about them and is the creator of their language. Producing books on HIV/Aids, farming techniques, malaria, etc. all help in the physical aspects of survival but also help their esteem when they can show others that they too have books in their language. Other aspects of what we do, developing dictionaries, grammars, etc. help document the language but more than that they benefit the language community. We have learned that instead of doing this as outsiders, if we engage the community in all of the above it empowers the community as a whole.



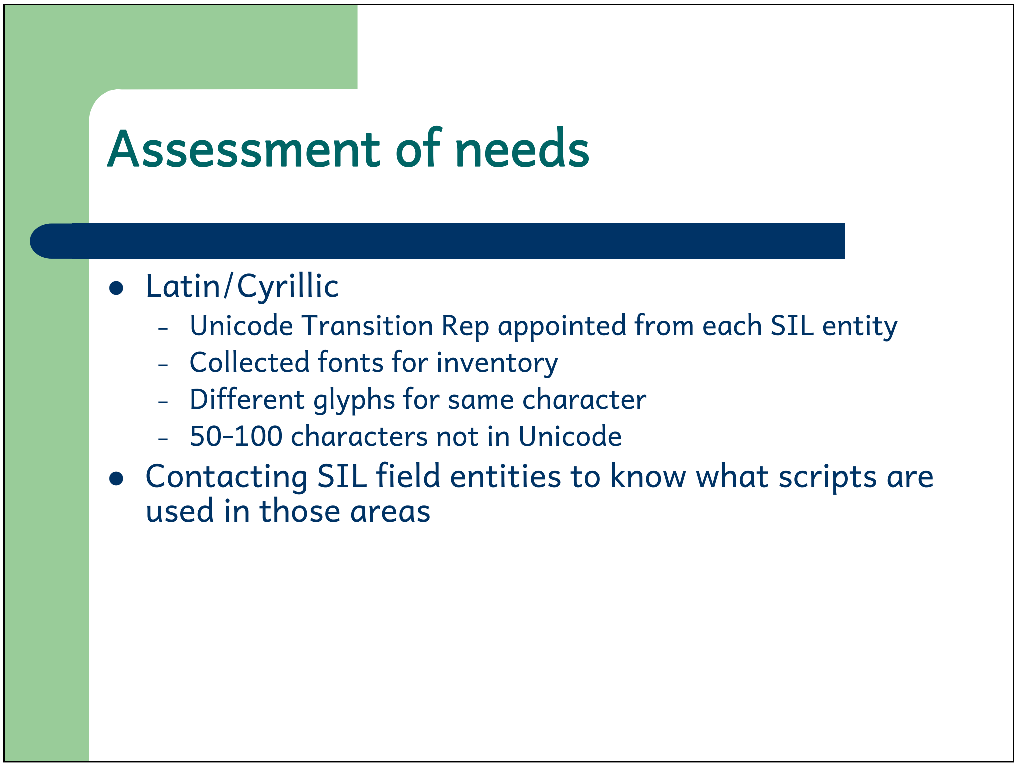
SIL’s contributions to endangered languages (cont.)

- Ethnologue – a catalogue of all known living languages, with language codes now merged with the international standard ISO 639-3
- Working with industry to enable support for these languages in software
  - OpenSource projects
  - Example: getting sample texts in +/- 600 languages for testing in a major application

SIL’s contributions to endangered languages (cont.)

Recent contributions in the IT world have included our 3-letter Ethnologue codes now merged with the international standard ISO 639-3 and working with industry and the OpenSource community to enable support for these languages in software.

Let’s look more specifically at the area of Unicode.



Assessment of needs

- Latin/Cyrillic
  - Unicode Transition Rep appointed from each SIL entity
  - Collected fonts for inventory
  - Different glyphs for same character
  - 50-100 characters not in Unicode
- Contacting SIL field entities to know what scripts are used in those areas

Assessment of needs

When we first decided to switch our organization to Unicode we knew it would be a big task to convince our linguists to switch to Unicode. Before we could even attempt that, Unicode fonts and Unicode applications had to be in place for them to use. My department, the Non-Roman Script Initiative (NRSI), was tasked with spearheading the transition of the organization to Unicode.

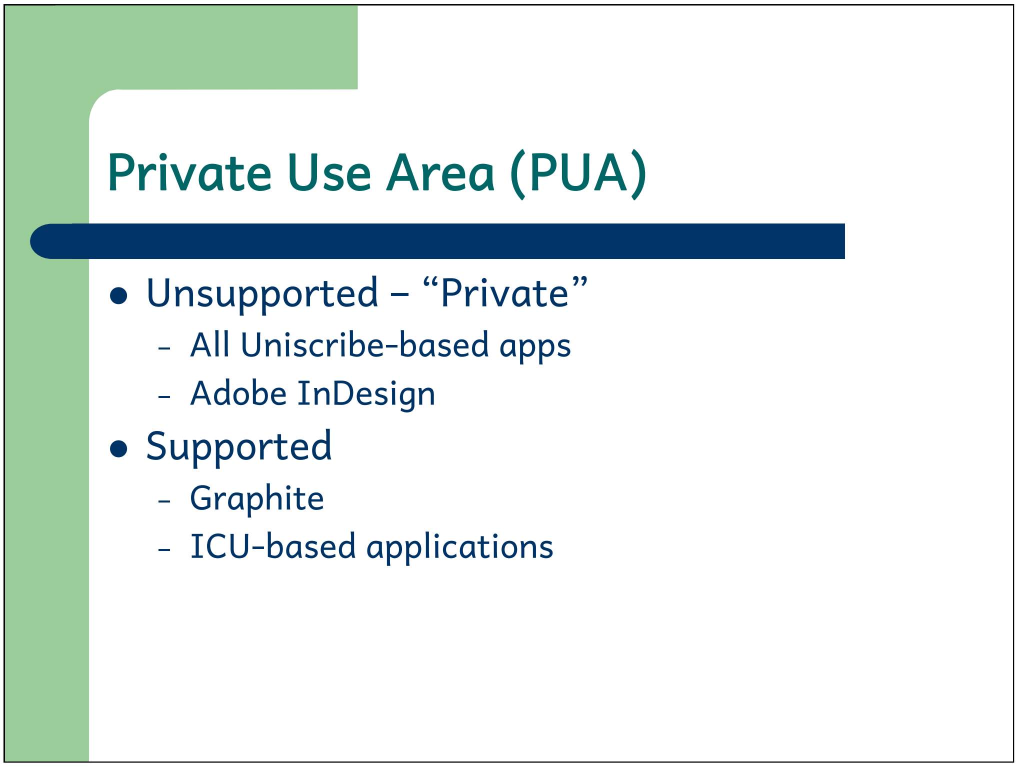
In assessing the needs we found that the writing-system needs in the Roman world was a big gray area. We didn't really have any idea how many legacy fonts (that is different encodings) were in use within our organization. The process of inventorying the character and glyph needs took approximately two years. We requested that each SIL entity appoint a Unicode Transition representative (UT representative). An SIL entity is basically a group of our linguists who are working in one geographical region. This Unicode Rep would be the one the NRSI would interact with. This person collected all of the fonts from their entity and sent them to the NRSI. With an in-house utility we were able to inventory all the glyphs in the fonts. Ultimately we had several hundred fonts we used for a glyph inventory. Once this inventory was in a database we could do a frequency count of glyphs, as well as knowing which glyphs mapped to which Unicode Scalar Value (USV). Through this process we found a number of alternate glyphs for the same character, and we also found between 50-100 characters which were not in Unicode. At this stage we went back to the UT representatives to find out if the character was in actual use. In some cases the character had been tested during orthography development and was never used. We did not include those in our fonts. Other times there was a definite need for that character.

Out of this process we came up with a list of characters which we put in SIL's corporate Private Use Area. These were included in our fonts and were the basis for making Unicode proposals in the Latin and Cyrillic ranges of Unicode.

Of course, Latin and Cyrillic were not our only needs. We contacted each of our field entities and asked them what non-Unicode-encoded scripts were in use in their areas.

A natural result of this assessment was knowing what characters and scripts need to be proposed for addition to Unicode.

This leads us to consider the Private Use Area (PUA).



Private Use Area (PUA)

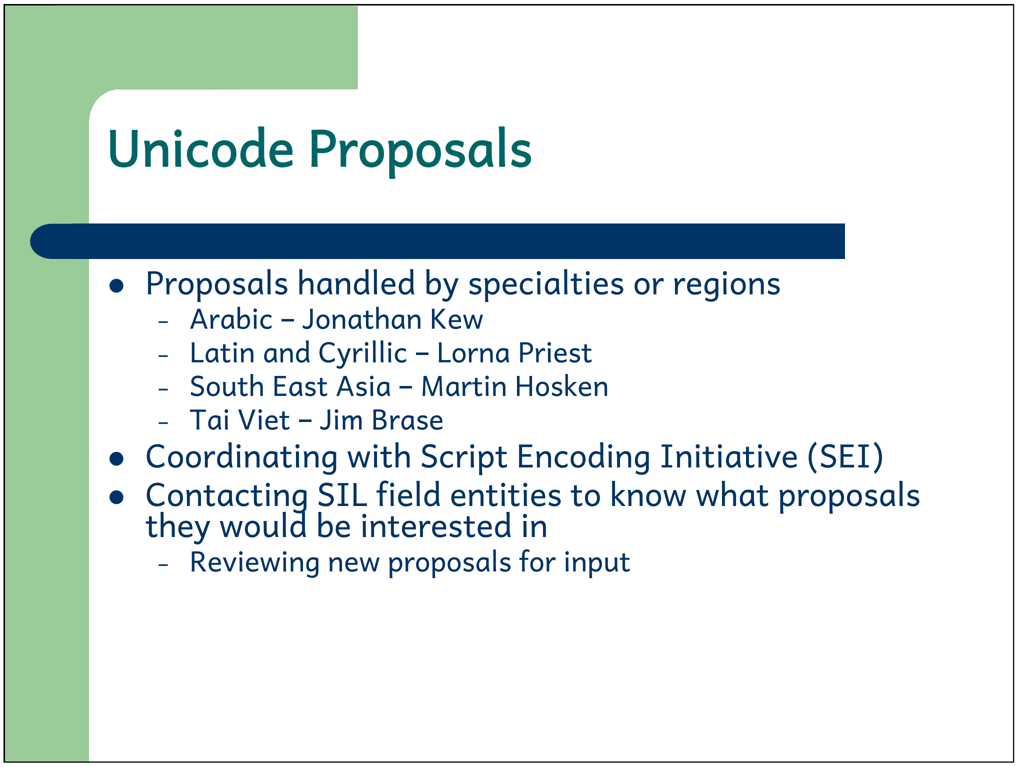
- Unsupported – “Private”
  - All Uniscribe-based apps
  - Adobe InDesign
- Supported
  - Graphite
  - ICU-based applications

Private Use Area (PUA)

In order for a proposal to be successful, Unicode expects the characters to be in use in published material. In order to be in use, the characters have to have a usable implementation for a number of years. In times past this was generally a custom encoding. In Unicode environments the only acceptable implementation involves PUA encoding. This is sufficient for only the simplest of scripts, but for anything beyond that smarts are needed. However, commercial vendors hesitate to implement smarts for characters in the PUA because that appears to be blessing a specific encoding in the PUA, and the PUA is for private use, not public use. The result is that users in developing countries, and specifically of interest endangered languages, get short-changed.

SIL’s Graphite easily handles the PUA, including any smart font rendering, but needs wider deployment. Also, ICU-based applications have a generic shaping engine that can use the PUA.

Having said that, the eventual goal is to make Unicode proposals so these scripts will be natively supported by Operating Systems and applications. We will look at this on the next page.

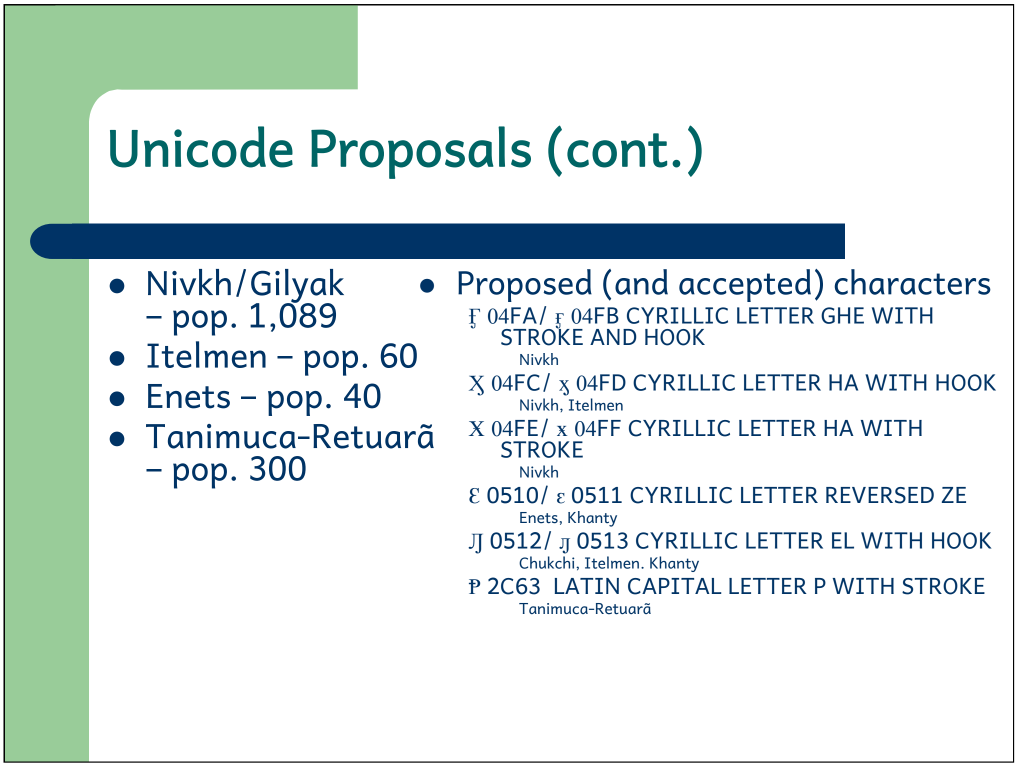


Unicode Proposals

- Proposals handled by specialties or regions
  - Arabic – Jonathan Kew
  - Latin and Cyrillic – Lorna Priest
  - South East Asia – Martin Hosken
  - Tai Viet – Jim Brase
- Coordinating with Script Encoding Initiative (SEI)
- Contacting SIL field entities to know what proposals they would be interested in
  - Reviewing new proposals for input

Unicode Proposals

Some types of Unicode proposals do not represent one particular area of the world. For example, Arabic and Cyrillic are used in many countries of the world, and Latin is used all over the world. Thus, we have not worked with a specific body of authority for those proposals. It has been sufficient to provide evidence of usage and need for those proposals. Other proposals require that we work with a local language authority or body. For example, for the Tai Viet proposal Jim Brase worked with an informal group called the Tai Viet working group. We often work with other people or organizations on proposals. Martin Hosken worked with Michael Everson on the Lanna proposal as well as the Burmese and Burmese extensions proposals. The Script Encoding Initiative has been generous in working with us. They have helped with funding and also with making sure there isn't overlap in proposals, that two proposals for the same script are not being worked on at the same time. If various people are interested in one script SEI has been helpful in bringing those parties together. We also keep in touch with SIL entities to know what proposals they may be interested in. If we know of a proposal in process then we can review it or send it to that field for review to make sure our needs are covered.



Unicode Proposals (cont.)

- Nivkh/Gilyak – pop. 1,089
- Itelmen – pop. 60
- Enets – pop. 40
- Tanimuca-Retuarã – pop. 300

- Proposed (and accepted) characters

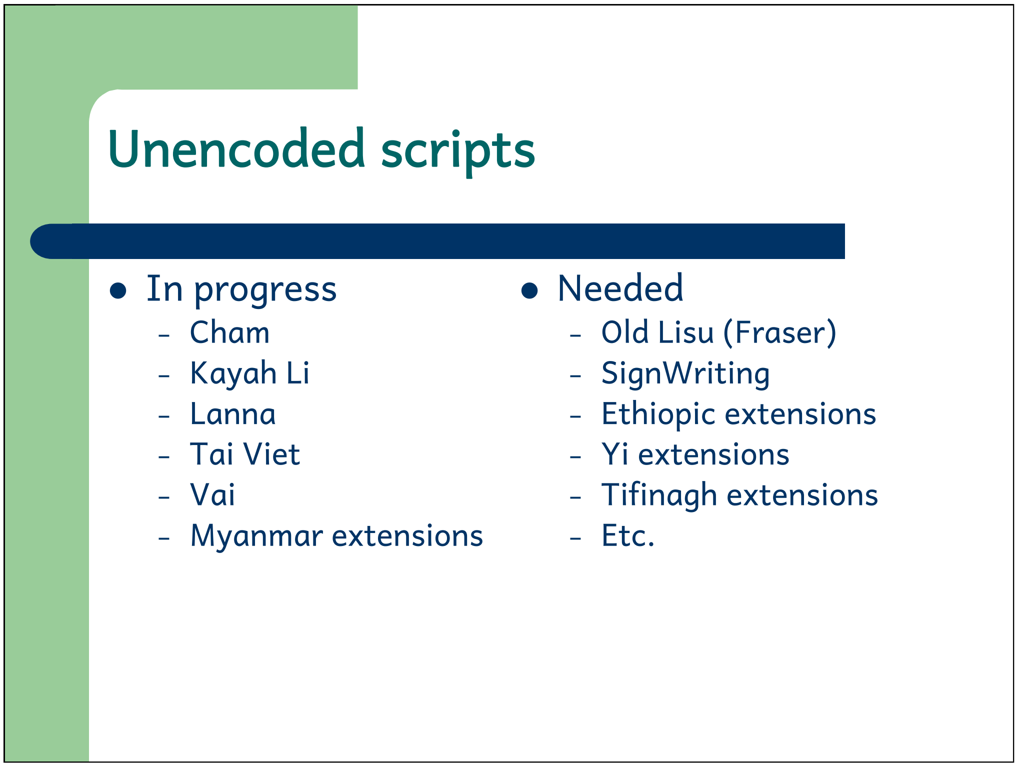
Ӻ 04FA/ ӻ 04FB CYRILLIC LETTER GHE WITH STROKE AND HOOK
Nivkh
Ӽ 04FC/ ӽ 04FD CYRILLIC LETTER HA WITH HOOK
Nivkh, Itelmen
Ӿ 04FE/ ӿ 04FF CYRILLIC LETTER HA WITH STROKE
Nivkh
Ԑ 0510/ ԑ 0511 CYRILLIC LETTER REVERSED ZE
Enets, Khanty
Ԓ 0512/ ԓ 0513 CYRILLIC LETTER EL WITH HOOK
Chukchi, Itelmen. Khanty
Ᵽ 2C63 LATIN CAPITAL LETTER P WITH STROKE
Tanimuca-Retuarã

Unicode Proposals (cont.)

We have proposed a number of characters which are in use by endangered languages. Examples of some of the groups are as follows: speakers of Nivkh number 1,089 (1989 census) although the ethnic population is 4,673, speakers of Itelmen number 60 although the ethnic population numbers 2,481 (1989 census), there are 40 speakers of Enets with a total ethnic population of 209 and the Tanimuca-Retuarã of Colombia number 300 (Ethnologue population figures). We proposed the following characters for these groups:

Ӻ 04FA/ ӻ 04FB CYRILLIC LETTER GHE WITH STROKE AND HOOK
Nivkh
Ӽ 04FC/ ӽ 04FD CYRILLIC LETTER HA WITH HOOK
Nivkh, Itelmen
Ӿ 04FE/ ӿ 04FF CYRILLIC LETTER HA WITH STROKE
Nivkh
Ԑ 0510/ ԑ 0511 CYRILLIC LETTER REVERSED ZE
Enets, Khanty
Ԓ 0512/ ԓ 0513 CYRILLIC LETTER EL WITH HOOK
Chukchi, Itelmen. Khanty
Ᵽ 2C63 LATIN CAPITAL LETTER P WITH STROKE
Tanimuca-Retuarã

The last pair of Cyrillic characters are used by the Khanty and Chukchi language communities as well as Itelmen. However, Khanty and Chukchi would probably not be considered endangered as they have 10,000 or more speakers.



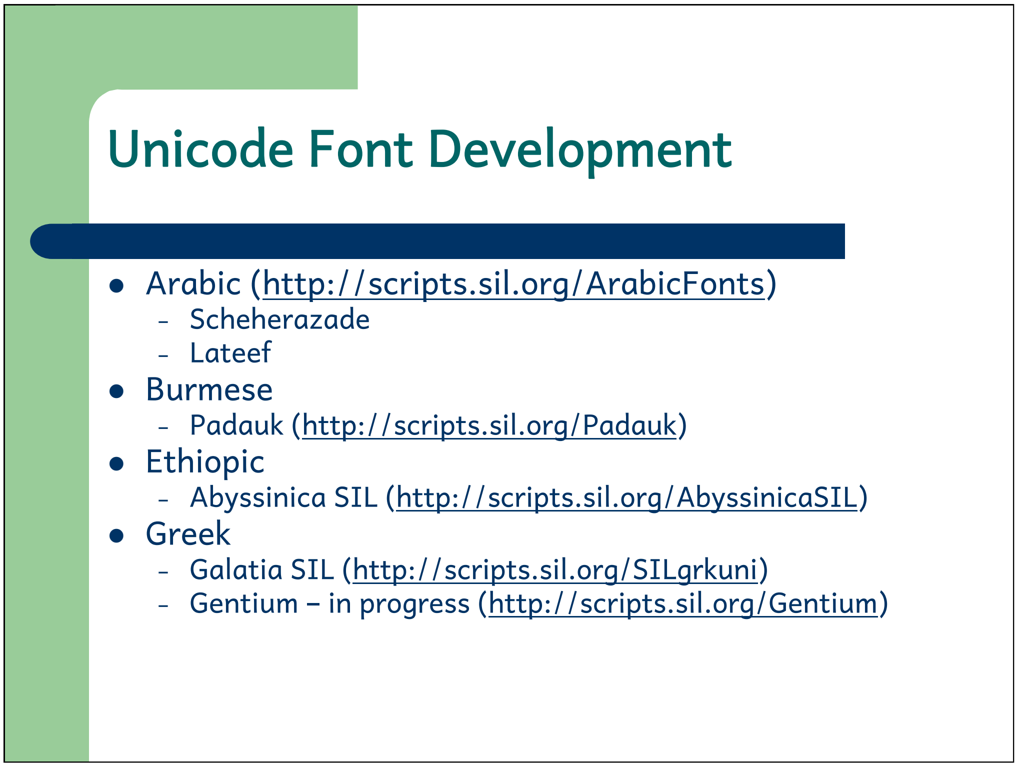
Unencoded scripts

- In progress
  - Cham
  - Kayah Li
  - Lanna
  - Tai Viet
  - Vai
  - Myanmar extensions
- Needed
  - Old Lisu (Fraser)
  - SignWriting
  - Ethiopic extensions
  - Yi extensions
  - Tifinagh extensions
  - Etc.

Unencoded scripts

The question is often asked: “Are there languages whose scripts are not yet encoded?” And the answer is “yes.” Some of the ones in which we are interested will be in Unicode 5.1 or subsequent versions of Unicode. We have been involved either through proposing, co-proposing, giving input or reviewing. Other proposals are in beginning stages of being written and others haven’t even begun to be written. There are many reasons for the slowness of proposals. Often the need is unclear, whether the script should be unified or disunified is a sticky issue, lack of personnel or funding to write proposals, and getting agreement from various interested parties is important and sometimes difficult.

Next, we’ll look at Unicode fonts and whether there are sufficient fonts out there.



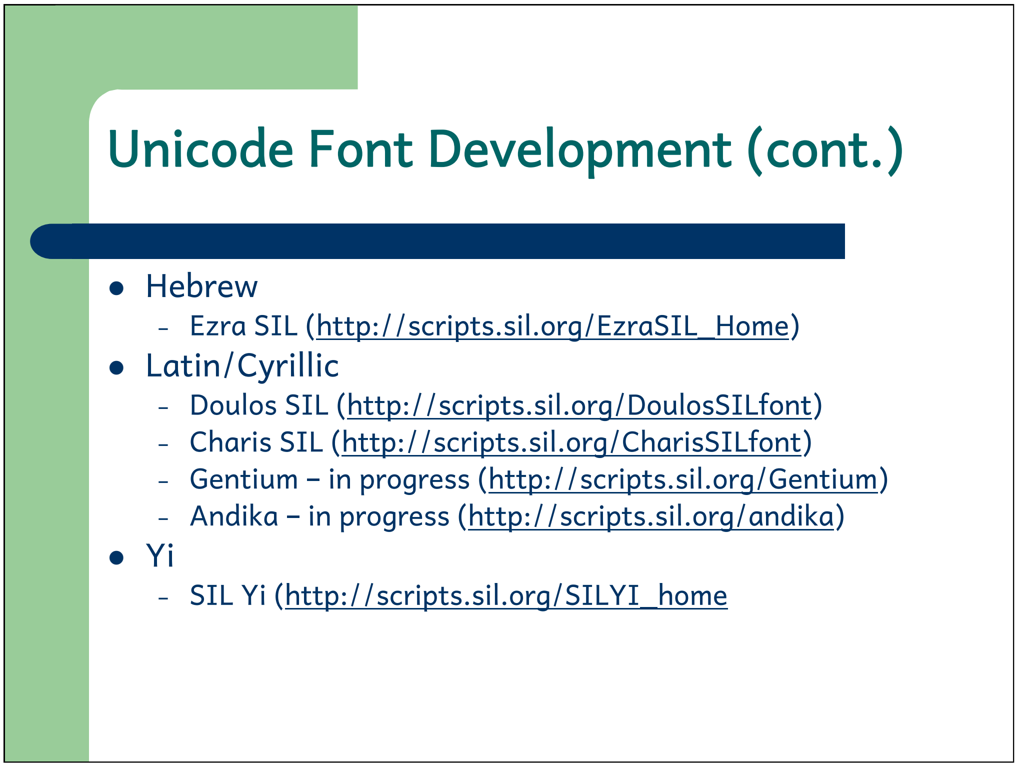
Unicode Font Development

- Arabic (http://scripts.sil.org/ArabicFonts)
  - Scheherazade
  - Lateef
- Burmese
  - Padauk (http://scripts.sil.org/Padauk)
- Ethiopic
  - Abyssinica SIL (http://scripts.sil.org/AbyssinicaSIL)
- Greek
  - Galatia SIL (http://scripts.sil.org/SILgrkuni)
  - Gentium – in progress (http://scripts.sil.org/Gentium)

Unicode Font Development

The fonts listed above, and on the next page, are all Unicode fonts which we have developed or had input in developing. Where needed, they include support for endangered languages. We have attempted to provide all the necessary support for any languages which use a particular script. However, we have not done specific research on the extent of how they may provide help for endangered languages.

When a person or organization makes a Unicode proposal they must also provide a font to go with that Unicode proposal. Padauk is an example of an existing Unicode font that was used for Unicode proposals.

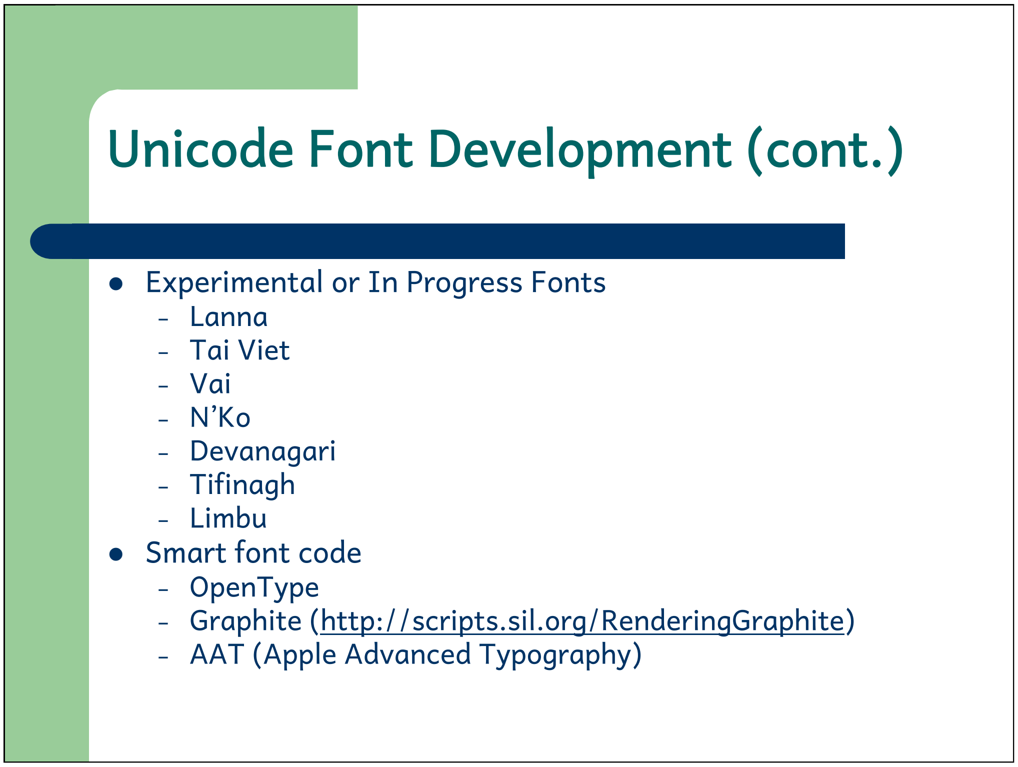


Unicode Font Development (cont.)

- Hebrew
  - Ezra SIL (http://scripts.sil.org/EzraSIL_Home)
- Latin/Cyrillic
  - Doulos SIL (http://scripts.sil.org/DoulosSILfont)
  - Charis SIL (http://scripts.sil.org/CharisSILfont)
  - Gentium – in progress (http://scripts.sil.org/Gentium)
  - Andika – in progress (http://scripts.sil.org/andika)
- Yi
  - SIL Yi (http://scripts.sil.org/SILYI_home

Unicode Font Development (cont.)

Most these fonts will be updated as need arises. Specifically, when new characters are added to Unicode or when new behavior is needed for a language.



Unicode Font Development (cont.)

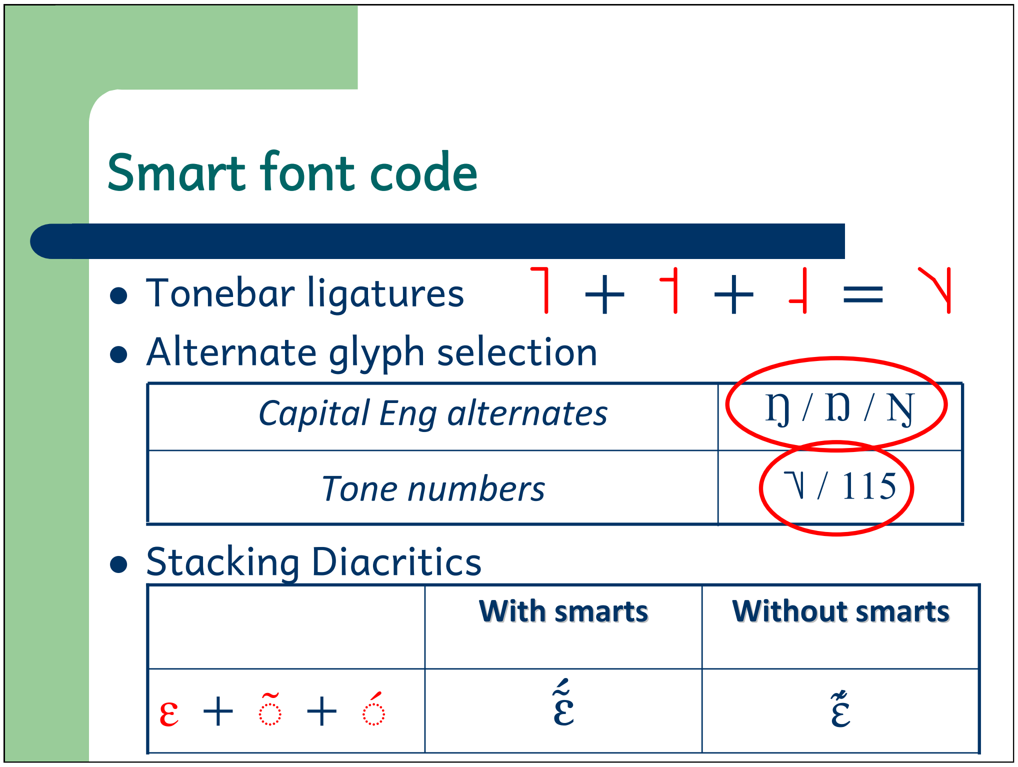
- Experimental or In Progress Fonts
  - Lanna
  - Tai Viet
  - Vai
  - N'Ko
  - Devanagari
  - Tifinagh
  - Limbu
- Smart font code
  - OpenType
  - Graphite (http://scripts.sil.org/RenderingGraphite)
  - AAT (Apple Advanced Typography)

Unicode Font Development (cont.)

For the Lanna and Tai Viet proposals, we are in the process of developing fonts. These will eventually be published as Unicode fonts once these scripts are an official part of the Unicode standard. We have also made our legacy font "SIL Vai" available under the Open Font License (more on that on slide 18). Work has already been done on turning this into a Unicode font (not by SIL).

We've assisted with Graphite code (http://scripts.sil.org/RenderingGraphite) in an N'Ko font. We are in the process of developing fonts for Devanagari, Tifinagh and Limbu. Through the process of the Tifinagh font development we are discovering that a further Unicode proposal will likely be needed for that script.

Most of our Unicode fonts require smart font code. Some contain code for all three rendering systems: OpenType, Graphite and AAT (Apple Advanced Typography). Some of the special features that are needed for specific languages can only be supported through Graphite.



Smart font code

- Tonebar ligatures ˥ + ˦ + ˨ = ˥˦˨
- Alternate glyph selection

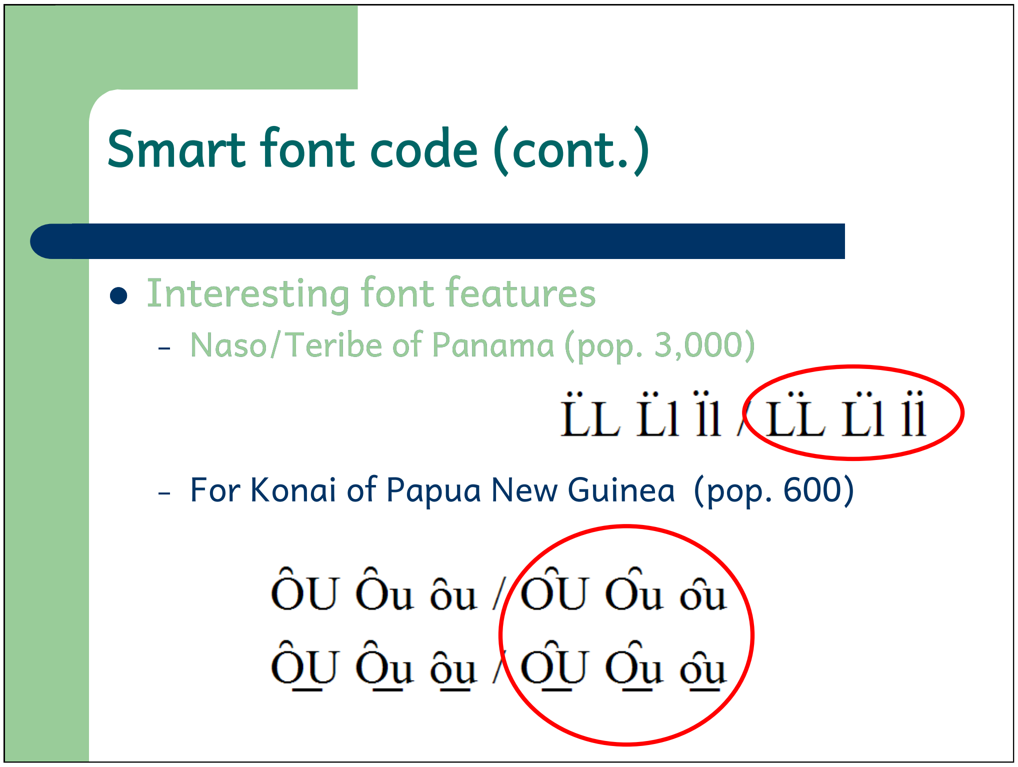
Capital Eng alternates	Ŋ / Ŋ / Ŋ
Tone numbers	˥˩ / 115

- Stacking Diacritics

	With smarts	**Without smarts**
ɛ + ◌̃ + ◌́	ɛ̃́	ɛ̃́

Smart font code

Some of you may be wondering what smart font code actually does. In the slide above we show three examples. The first one illustrates the use of ligatures. Ligatures are not only used in IPA with tonebars they are also used in many scripts. The second one illustrates the use of alternate glyph selection. In the first example, there are at least three different variations of the eng in use in different languages. In the second example we see that tone numbers are alternates to tonebars. At the moment, only Graphite applications and Adobe InDesign can use alternate glyph selection. A third example is how diacritics can stack on top of the base character. This feature is also needed in a number of South-east Asian scripts such as Thai, Lao, and Lanna.



Smart font code (cont.)

- Interesting font features
  - Naso/Teribe of Panama (pop. 3,000)

    L̈L L̈l l̈l / L̈L L̈l l̈l

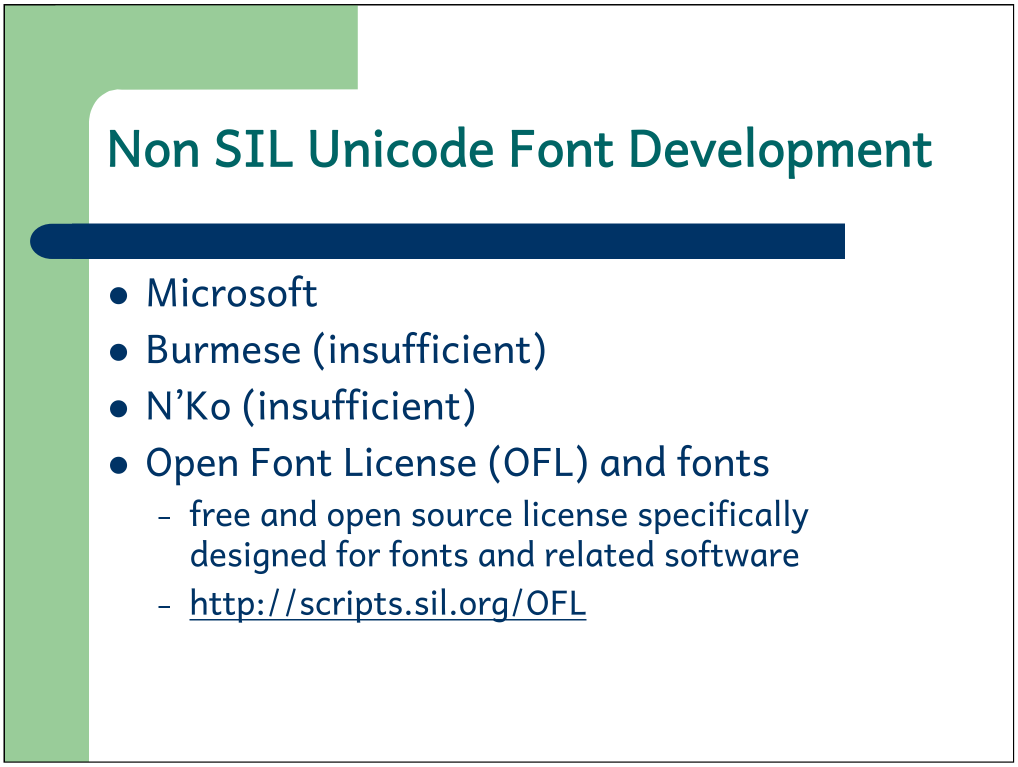
  - For Konai of Papua New Guinea (pop. 600)

    ÔU Ôu ôu / O͡U O͡u o͡u

    Ô̲U̲ Ô̲u̲ ô̲u̲ / O̲͡U̲ O̲͡u̲ o̲͡u̲

Smart font code (cont.)

A number of very small language groups had need for some very interesting font features. The Naso/Teribe people of Panama number only 3,000 people. However, they have an orthography which has been sanctioned by their king and also by Panama which utilizes two dots centered over the double-el digraph. In the same way, the Konai people of Papua New Guinea (population 600) have the inverted breve which is centered over an ou diagraph. When we made a Unicode proposal for Naso, UTC deemed this was just a dieresis and not a double diacritic and should be handled with smart fonts. We have added these into our Latin/Cyrillic fonts. However, at this time, the only rendering engine that can handle them is Graphite. The Graphite code enables them to turn on this feature for their language.



Non SIL Unicode Font Development

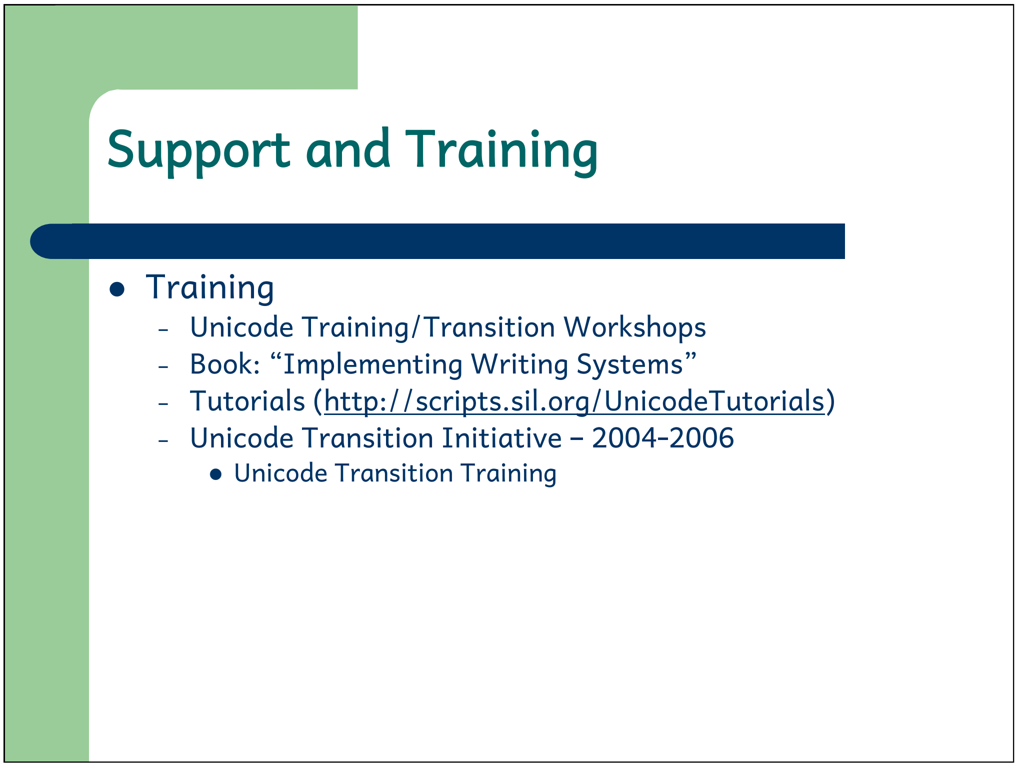
- Microsoft
- Burmese (insufficient)
- N’Ko (insufficient)
- Open Font License (OFL) and fonts
  - free and open source license specifically designed for fonts and related software
  - http://scripts.sil.org/OFL

Non SIL Unicode Font Development

With Microsoft’s recent Vista release there are many Unicode fonts available for many of the scripts already in Unicode. Of course, these are not freely distributable to those not using Vista, which provides limited availability for many of the language groups we work with. They also do not provide some of the features that smaller language groups may require.

Not only are fonts unavailable but commercial rendering engines may not yet be up to the task for some scripts. An example of this is the Burmese script. The only available Unicode solution at this time is the Padauk font. While Uniscribe does not support Burmese, some enterprising people have managed to get a Burmese font working using basic Uniscribe shaping based on the Padauk font. There is also a font being developed by the Myanmar NLP Community that uses the same approach. The best support at the moment is provided by Padauk using Graphite. Another example is the N’Ko script. This is a right-to-left script and although there is a Graphite font available the only software up to the task of rendering it are XeTeX (http://scripts.sil.org/XeTeX) and SIL FieldWorks (http://www.sil.org/computing/fieldworks/). Efforts are under way to make OpenOffice render N’Ko.

In partial answer to the problem of insufficient fonts, we have created, in conjunction with the Open Source community, a new license called the Open Font License (OFL). We are now distributing all of our new fonts under the Open Font License (OFL). We believe in making software and fonts available for everyone to use, in particular those who may have no representation otherwise. We hope this will encourage others to open up their fonts for free use and also for development so that if a font does not provide the support needed, the license would allow for further development.



Support and Training

- Training
  - Unicode Training/Transition Workshops
  - Book: “Implementing Writing Systems”
  - Tutorials (http://scripts.sil.org/UnicodeTutorials)
  - Unicode Transition Initiative – 2004-2006
    - Unicode Transition Training

Support and Training

One last issue was to address the need for training. We have organized a number of Unicode Transition or Unicode Training workshops. Attending these workshops were people who needed to understand Unicode for their work as well as people whose job is to do computer support and/or training.

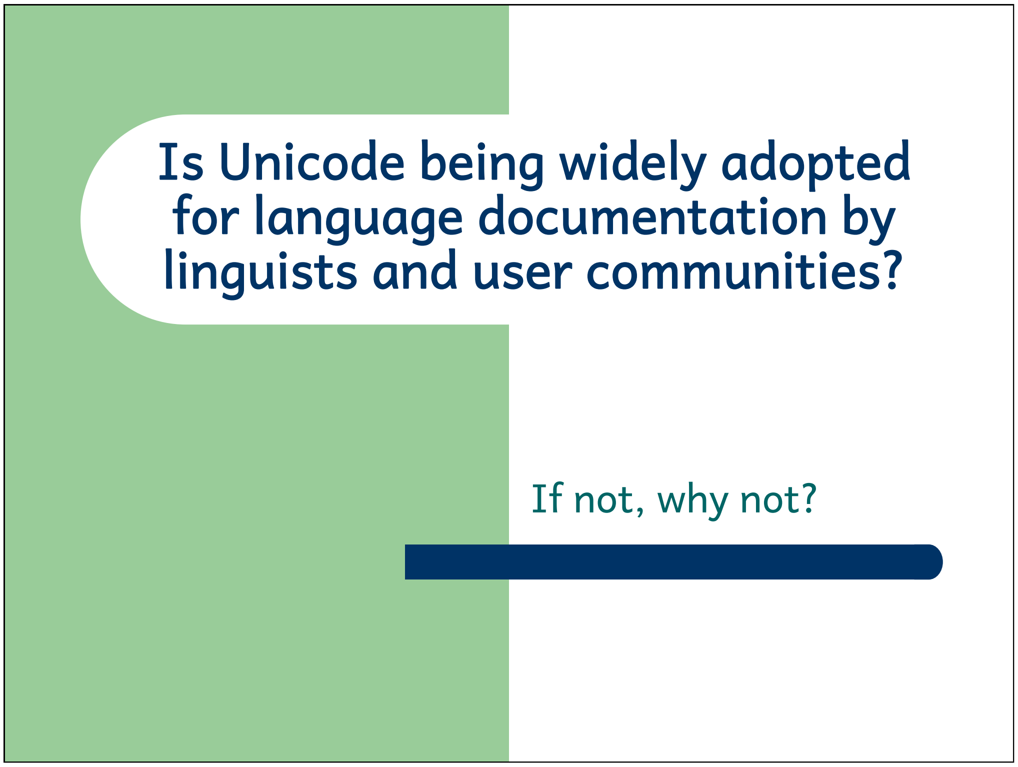
In the process of pulling together the agenda for the first workshop we realized the need for a resource handbook to give the participants. Out of this came a book called “Implementing Writing Systems” (http://scripts.sil.org/IWS-TOC).

Also out of these workshops came some tutorials that we developed and made available online (http://scripts.sil.org/UnicodeTutorials). These tutorials are quite technical.

Because of the slowness in adopting to Unicode, our administration came to the conclusion that transitioning to Unicode should be a main initiative for a two-year period. Successful completion would include fonts, tools, training and implementation. The training portion brought together individuals from various academic domains in SIL: linguistics, anthropology, literacy, etc. All the training prior to this was directed at computer support people, not the ordinary linguist. Modules were developed using Moodle, an OpenSource course management system (http://moodle.org/). These courses have proven useful; however, the login requirement for taking a course has been a deterrent for some. Training has proved to be the most difficult for us, and at the end of the two-year period we concluded that development of training materials was only partially completed. We anticipate that development of training materials should continue. However, now that the “initiative” is over, personnel are no longer officially assigned to this task and we are finding they no longer have time to commit to this.

We have seen many people come out of the workshops ready to use Unicode for their own data. “Unicode evangelists” have come from the workshops. We’ve seen where tools, fonts and training needed to be strengthened in order for users to successfully switch to Unicode. And the job *has* been very challenging. Nothing is straightforward when you attempt to convert legacy data to Unicode (this paper does not address any of those issues).

Now, let’s look at the next question...



Is Unicode being widely adopted for language documentation by linguists and user communities?

If not, why not?

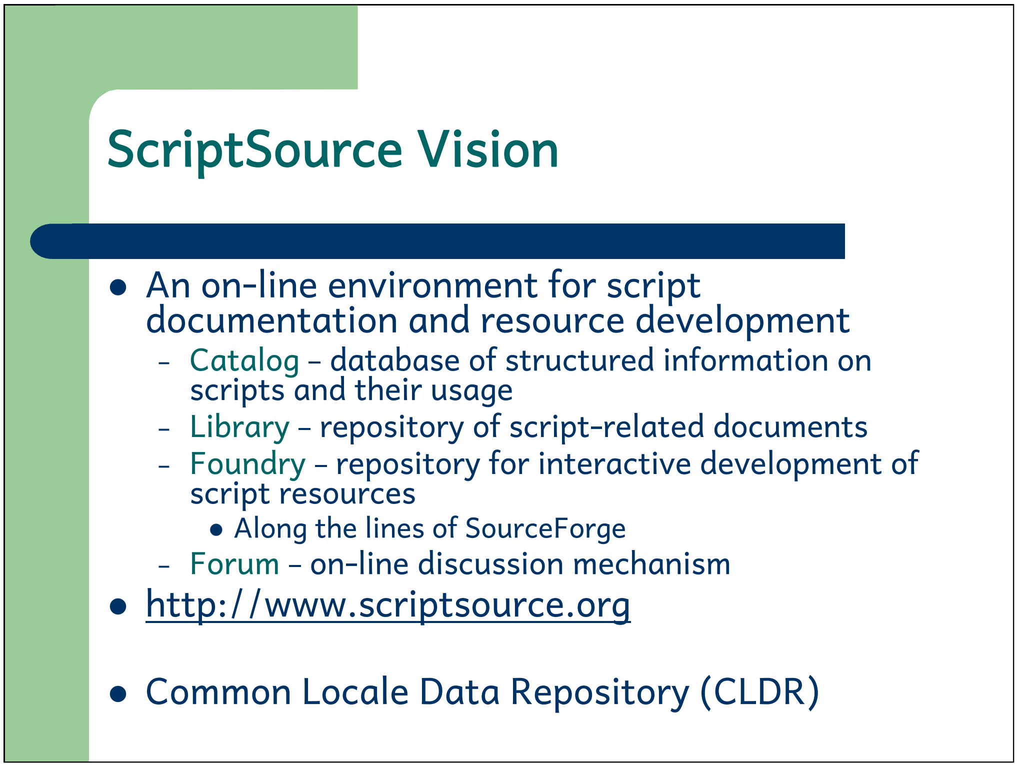
Is Unicode being widely adopted for language documentation by linguists and user communities?

I believe it is being widely adopted. Each semester at the Graduate Institute of Applied Linguistics my department (Non-Roman Script Initiative) gives a talk to students about Unicode and the difficult issues with other writing systems, keyboarding, fonts, smart fonts... The last time we did this, the response from one student was “I didn’t know it could be so difficult, everything has been very straightforward for me.” This response made us more aware that we have been successful! Now students *begin* with using Unicode, they don’t have to convert from using legacy fonts and keyboards and it is not so complicated. When new projects can begin with Unicode the task isn’t so daunting.

However...

At a lunch-time conversation with a professor I was asking if he used Unicode. He said, “no, why should I?” I was rather taken aback. He is not in the linguistic domain, doesn’t need IPA very much, the fonts he uses have everything he needs, he doesn’t ever send his computer files to anyone else to use, always gives his students printed handouts and sees no need. What could I say?

When you look beyond the classroom and into the real world of linguistics, there are a number of existing projects where Unicode has not been adopted or not been adopted completely. We still have people with older computers that cannot run Unicode-enabled applications. Sometimes we have a team of workers on one project where some have new computers and others have old computers. They still need to share their data. Some teams don’t have good Unicode font solutions. Because of these issues, we see teams who are converting their data back and forth from legacy encodings to Unicode. As this appears to be a successful round-tripping, they feel that when everything is in place (applications and fonts and computers) then it will be easy to convert to Unicode when they are ready.



ScriptSource Vision

- An on-line environment for script documentation and resource development
  - Catalog – database of structured information on scripts and their usage
  - Library – repository of script-related documents
  - Foundry – repository for interactive development of script resources
    - Along the lines of SourceForge
  - Forum – on-line discussion mechanism
- http://www.scriptsource.org
- Common Locale Data Repository (CLDR)

ScriptSource Vision

A project that we feel will help in this whole process of using Unicode and writing system support for individual languages is ScriptSource. We anticipate having an on-line environment for script documentation and resource development. It would include:

Catalog: database of structured information on scripts and their usage

Library: repository of script-related documents

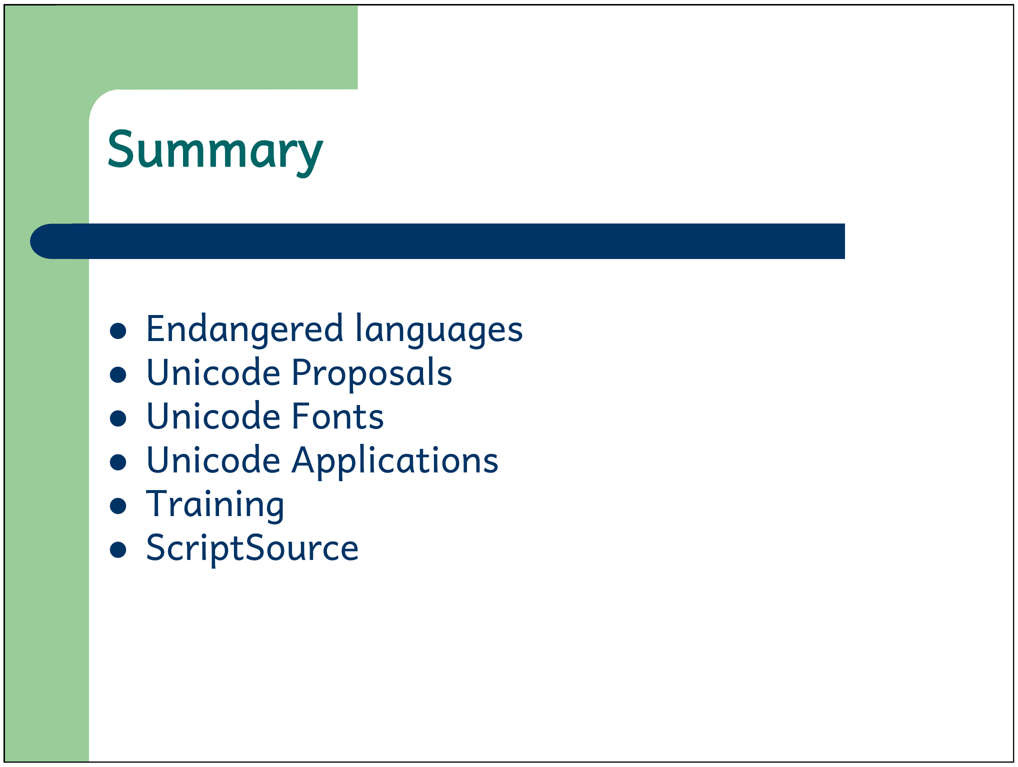
Foundry: repository for interactive development of script resources

Along the lines of SourceForge

Forum: on-line discussion mechanism

We believe that ScriptSource would help “safe” *and* endangered languages alike. If the writing system information is in the repository then when an application developer wanted to support any language, they would be able to go to ScriptSource to find out what is needed for implementing support for that language. Developers will also be able to use ScriptSource to collaborate on projects together, sharing resource and knowledge.

A similar collaborative project (although the scope is much smaller) is the Common Locale Data Repository (CLDR). The question was asked “Is there a way to encourage more locale data submission?” Up to this point, we have not been involved in this. Since we do not currently have the personnel to follow up on this and encourage linguists in this area we have not been working with linguists to submit data to the CLDR. However, if we were to incorporate submission of this type of data into ScriptSource it would be natural to make sure the information was also submitted to the CLDR.



Summary

- Endangered languages
- Unicode Proposals
- Unicode Fonts
- Unicode Applications
- Training
- ScriptSource

Summary

As we have seen, slowing or stopping the demise of a language is something we should all be interested in. Where we cannot stop the loss of a language, at least we can document it so as to not lose what we may learn from that language. Unicode can help us in this regard by making sure that language can be represented in this industry standard.

We've also seen that the task of encoding scripts is not finished. More are waiting. In some cases there are people willing to make proposals; funding is lacking for the research necessary to understand the script and to get consensus from all interested parties.

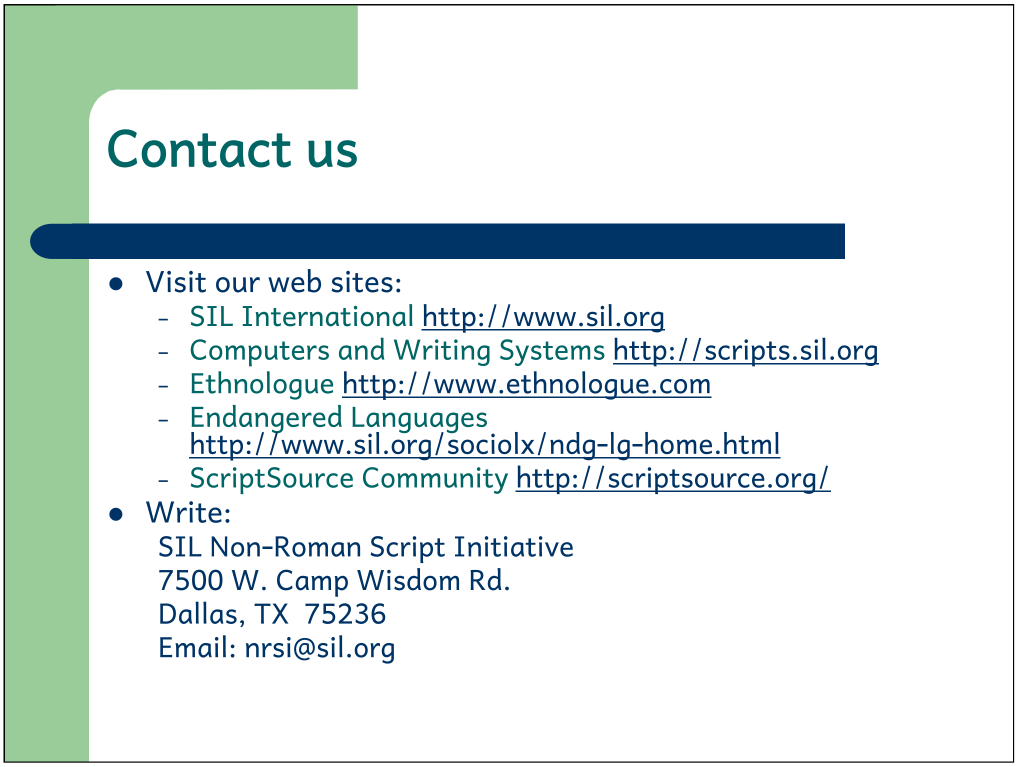
We've come a long way in the last few years in having fonts available which can render text in these languages. There are still great needs, especially for fonts which meet the needs of smaller user communities. We have great hopes that our Open Font License will be used by many to make fonts freely available. Already there are a number of non-SIL fonts being released under this license.

Although it has not been a major topic of this paper, it is clear that applications are lacking. Rendering engines may not support the writing system behavior. Applications may not be available for rendering systems which do support the writing system behavior needed by these smaller languages.

Training is still an area of need. Instead of requiring users to take time out for training we need more targeted modules to help the user in the area of his or her immediate problem.

We believe ScriptSource will be a huge help for application and font developers to understand the behavior that is need for a particular language and writing system, as well as for linguists, academics, standards bodies and governments.

I hope this paper has been helpful in seeing a little of the state of Unicode in relation to smaller language communities such as endangered languages.



Contact us

- Visit our web sites:
  - SIL International http://www.sil.org
  - Computers and Writing Systems http://scripts.sil.org
  - Ethnologue http://www.ethnologue.com
  - Endangered Languages http://www.sil.org/sociolx/ndg-lg-home.html
  - ScriptSource Community http://scriptsource.org/
- Write:

  SIL Non-Roman Script Initiative
  7500 W. Camp Wisdom Rd.
  Dallas, TX 75236
  Email: nrsi@sil.org

Bibliography

Cahill, Michael. 2004. *From Endangered to Less Endangered: Case Histories from Brazil and Papua New Guinea*. SIL Electronic Working Papers. http://www.sil.org/silewp/2004/silewp2004-004.htm

Crystal, David. 2000. *Language Death*. Cambridge: Cambridge University Press.

Gordon, Raymond G., Jr. (ed.), 2005. *Ethnologue: Languages of the World*, Fifteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

Headland, Thomas N. 2003. *Thirty Endangered Languages in the Philippines*. Work papers of the Summer Institute of Linguistics, University of North Dakota Session, vol. 47 (2003) http://www.und.nodak.edu/dept/linguistics/wp/2003Headland.PDF

Krauss, Michael. 1992. *The world's languages in crisis*. Language: Journal of the Linguistic Society of America 68:4-10. Published by the Linguistic Society of America at the Waverly Press Inc., Baltimore, MD. 21202.

UNESCO. 2005. *Promoting the Convention for the Safeguarding of the Intangible Cultural Heritage: (information kit)*. http://unesdoc.unesco.org/images/0014/001412/141247e.pdf