Software Globalization and Adding Languages on Computers and Mobile Devices

Craig Cummings

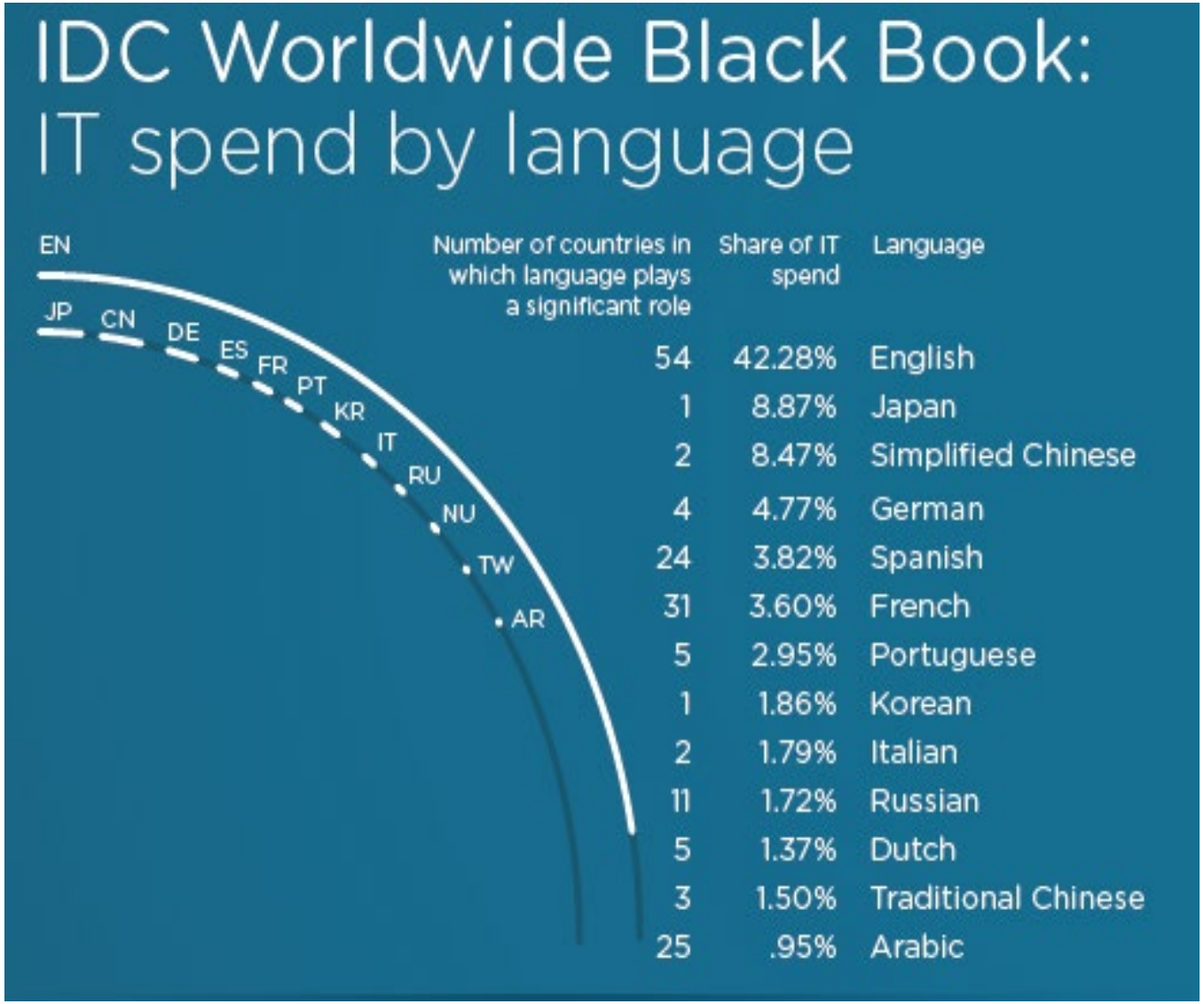
Unicode Technical Committee Vice-Chair

ANSI INCITS L2 Committee Chair

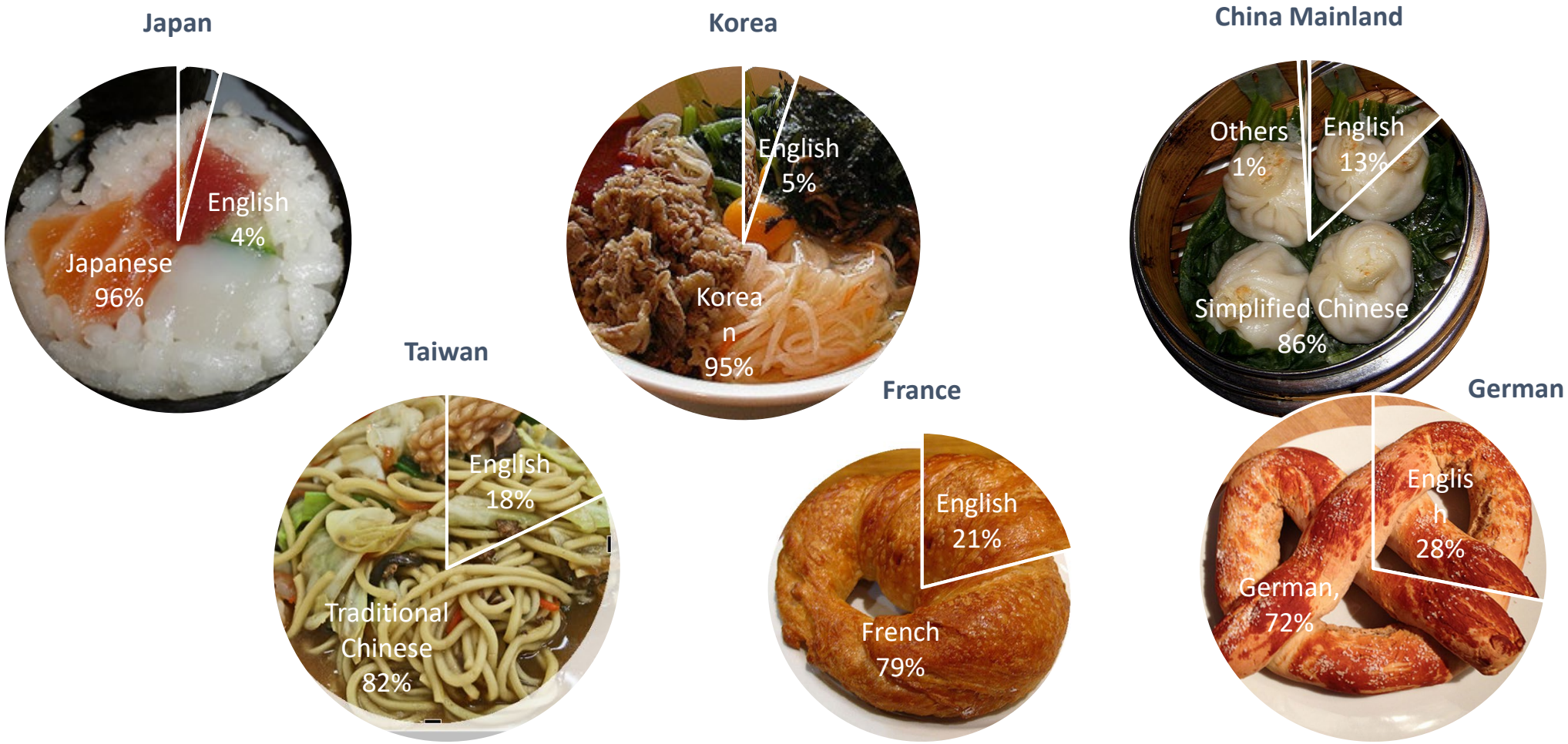
Unicode Emoji Subcommittee Animal/Science Category Manager

VMware Globalization Team Lead

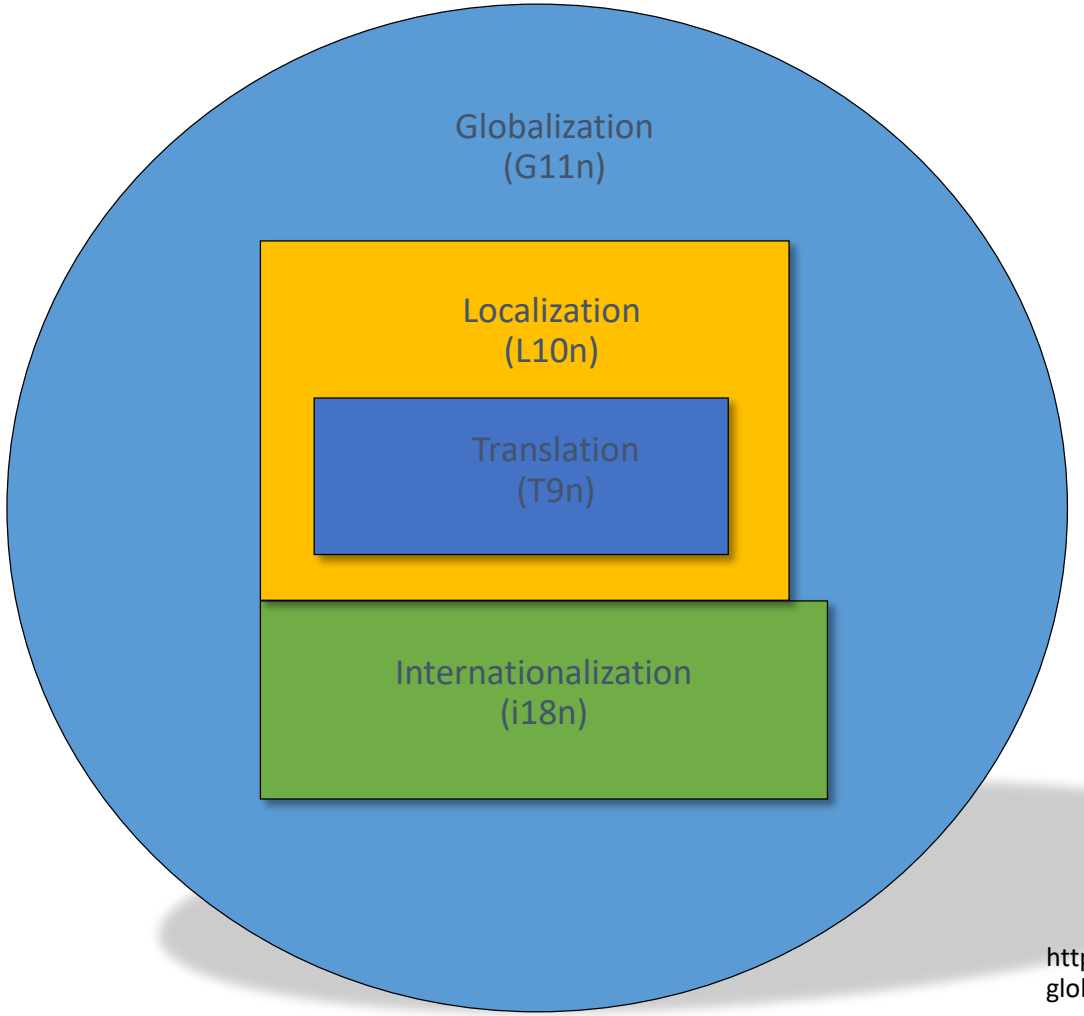
Why Globalize Software?



Why Globalize Software? -- English vs. Localized Page Views



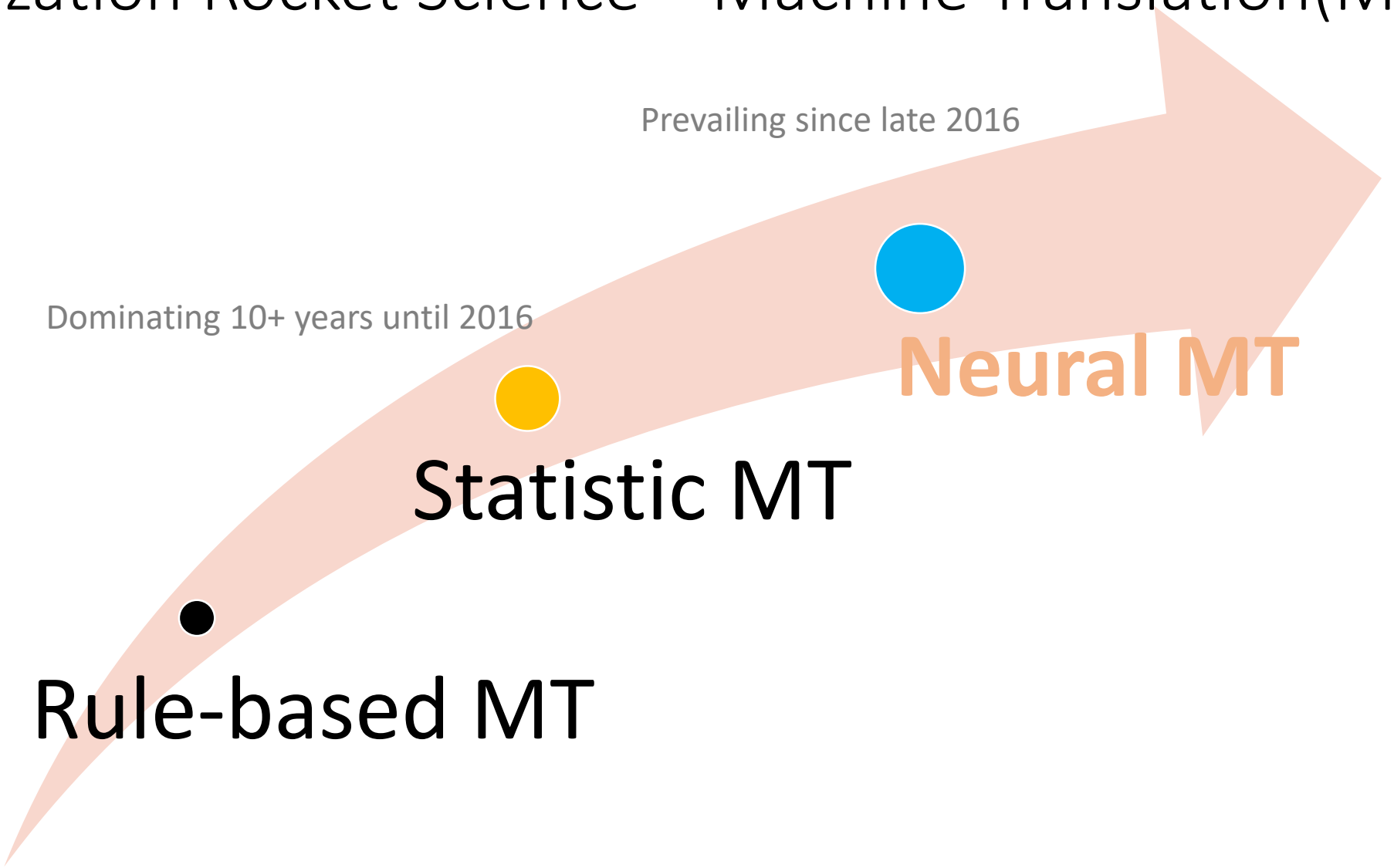
G11n, i18n, L10n and T9n



- Globalization (G11n): The complete process of making your application available in multiple languages consisting of 2 sub processes Localization and Internationalization.
- Internationalization (i18n): Architecting and coding for language/locale independence; allows localization for target audiences that vary in culture, region, or language.
- Localization (L10n): Adapting to specific languages/locales by translating text and adding locale-specific formatting to images and style sheets.
- Translation (T9n): Converting the meaning of text in one language into another.

https://blogs.adobe.com/globalization/2012/04/12/globalization-myth-series-myth-1-software-globalization-internationalization-localization-translation/

Localization Rocket Science – Machine Translation(MT)



Animation source: https://techcrunch.com/2017/04/13/neural-networks-made-easy/

java.io.InputStreamReader

(use the constructor that includes an explicit encoding)

Code this 

```
new InputStreamReader(outputStream, explicitEncoding);
```

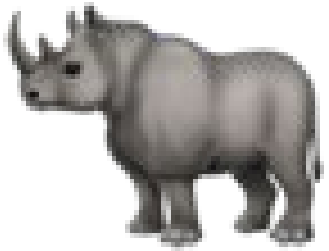
Not this

```
new InputStreamReader(outputStream);
```

Internationalization Eye Chart

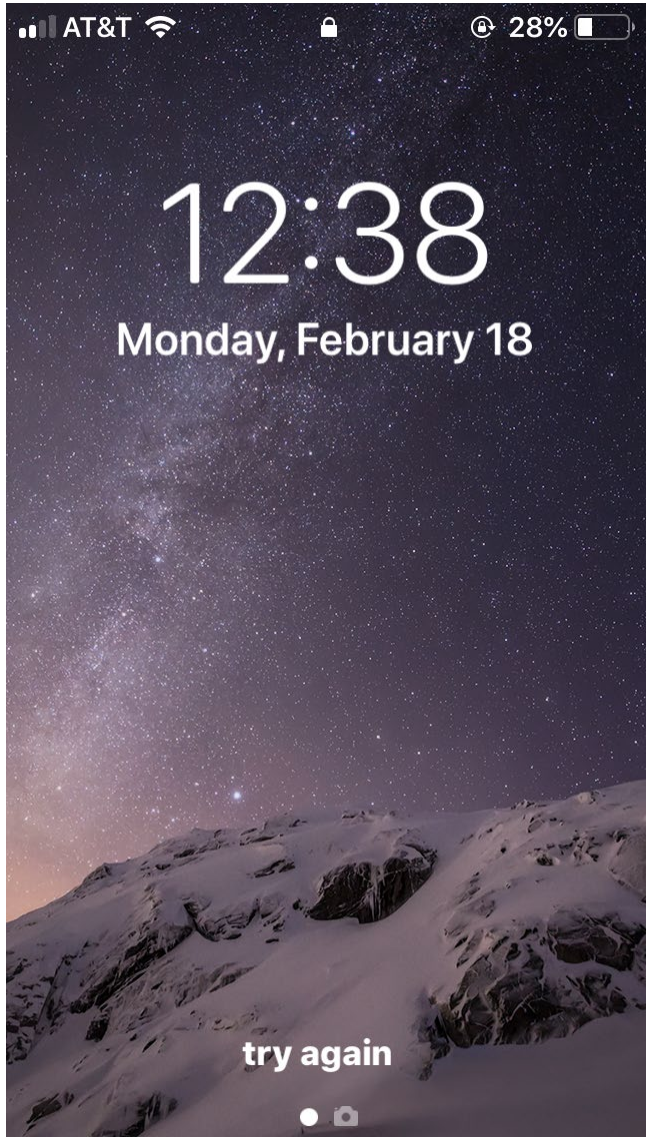
- Encodings, I/O, character support
  - Regular expressions including Unicode regular expressions
    - Because [A-Za-z0-9] works only for ASCII
  - Internationalized Domain Names (IDNs)/Internationalized Resource Identifiers (IRIs)
  - Supplementary character support
- Locales (BCP-47 compliant)
  - Determination
  - Representation
  - Negotiation
  - Fallback
- Locale-sensitive APIs
  - Externalization
    - Resource files and messages
  - Formatting
    - Dates, times, time zones and calendars
    - Numbers
    - Currencies
    - Honorifics
    - Addresses and phone numbers
  - Linguistic (not binary) Sorting
  - Wrapping/boundary analysis (words, lines, sentences, etc.)
- Keyboard support
- User Interface (including bidi), fonts, and layout support
- Search, normalization, folding, etc.
- Security
- …and more

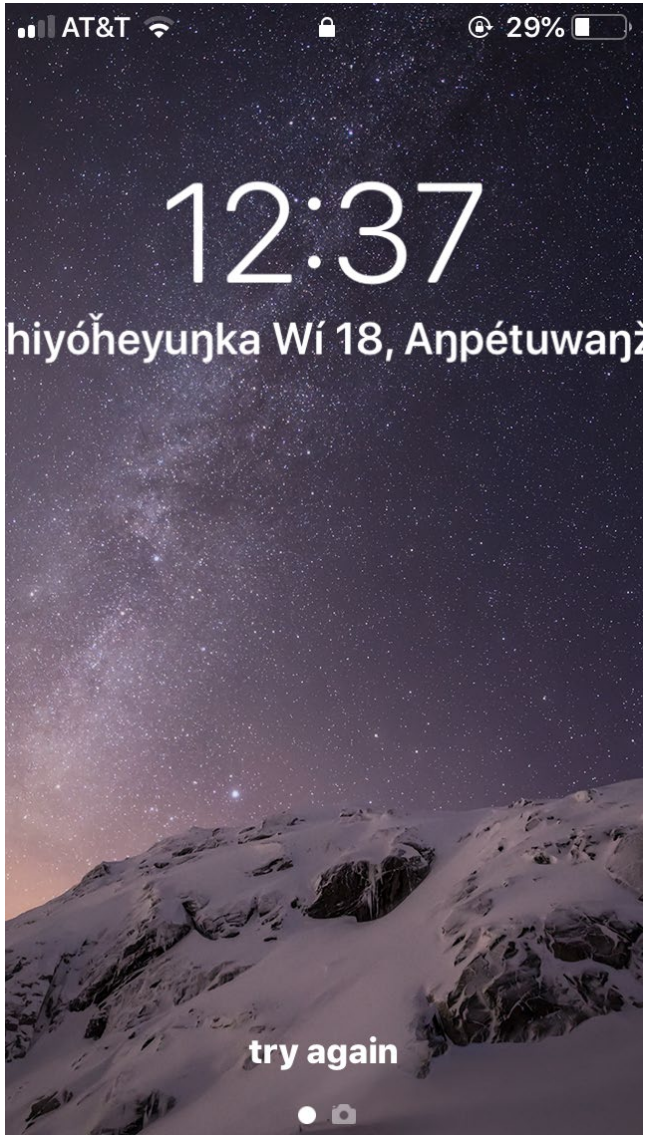
You Probably Know Unicode for…

Hoping you know Unicode for writing systems

Did you know Unicode for this, too?





The Three (minimal) Requirements To Get Your Language On Computers and Mobile Devices

1. Fonts (already there for Latin)
2. Keyboards (layouts now a part of CLDR)
3. Unicode CLDR data (the short straw)

CLDR = Common Locale Data Repository

CLDR - Unicode Common Locale Data Repository

Navigation

Unicode CLDR Project
CLDR Releases/Downloads
CLDR Survey Tool
CLDR Change Requests
CLDR Charts
CLDR Process
CLDR Specifications
Information Hub for Linguists
Unicode Extensions for BCP 47
Implementer's FAQ
ULI Subcommittee

Milestone Schedule

Q2/3		Targets
Apr	01	Start Tool/Data Preparation
May	16	Start **Shakedown Submission**
May	23	Start **General Submission**
		Start **Limited Submission**
Jul	11	End Submission Start **Vetting**
Jul	25	End Vetting **Start Resolution**
Aug	08	Start **Production**
Aug	22	**Data Freeze** All manual data changes done, only BRS data changes thereafter
Sep	12	**Alpha — Final Data candidate** No change to data affecting ICU thereafter Other dtd, data, spec, docs, tool changes allowed
Sep	26	**Beta — Final Candidate** No dtd or data changes allowed thereafter Docs, charts, spec changes allowed (= ICU release candidate)
Oct	15	**Release**

Unicode CLDR Project > CLDR Specifications >

Core Data for New Locales

This document describes the minimal data needed for a new locale. There are two kinds of data:

1. Core XML Data - This is data that the CLDR committee needs from the proposer before a new locale is added. The proposer is expected to also get a Survey Tool account, and contribute towards the Minimal Data.
2. Minimal Data Commitment - Data that is expected to be provided for each locale. If it is not supplied in a timely fashion, the committee may remove the locale.

(The parenthesis at the start of each line below has the approximate number of strings for each item.)

Core XML Data

First, make sure you have correct language code according to Picking the Right Language Identifier. Then collect the following data. Consider using the Core Data Submission Form to submit this data.

Note to translators: If you are having difficulties or questions about the following data, please contact us. Post a follow-up to your existing bug, file a new bug, or reply to the mailing list.

1. (04) Exemplar sets: main, auxiliary, index. **[main/xxx.xml]**
   - These must reflect the Unicode model. For more information, see tr35-general.html#Character_Elements.
2. (02) Orientation (bidi writing systems only) **[main/xxx.xml]**
3. (01) Plural rules **[supplemental/plurals.xml]**
   - For more information, see cldr-spec/plural-rules.
4. (01) Default content script and region (normally: normally country with largest population using that language, and normal script for that). **[supplemental/supplementalMetadata.xml]**
5. (N) Verify the country data (i.e. which territories in which the language is spoken enough to create a locale) **[supplemental/supplementalData.xml]**
6. (N) Casing information (cased scripts only, according to ScriptMetadata.txt)
   - This will be in common/casing
7. (N) Collation rules [non-Survey Tool]
   - For details, see cldr-spec/collation-guidelines.
   - The result will be a file like: common/collation/ar.xml or common/collation/da.xml.
   - Note that the "search" collators (which tend to be large) are not needed initially.

Recommended Core Data

The following are not required, but are strongly recommended:

1. (04) Exemplar set: punctuation. **[main/xxx.xml]**
2. (01) Ordinal rules **[supplemental/ordinals.xml]**
   - For more information, see cldr-spec/plural-rules.
3. *(N) Romanization table (non-Latin writing systems only) **[spreadsheet, we'll translate into transforms/xxx-en.xml]**
   - If a spreadsheet, for each letter (or sequence) in the exemplars, what is the corresponding Latin letter (or sequence).
   - More sophisticated users can do a better job, supplying a file of rules like transforms/Arabic-Latin-BGN.xml.

Minimal Data Commitment

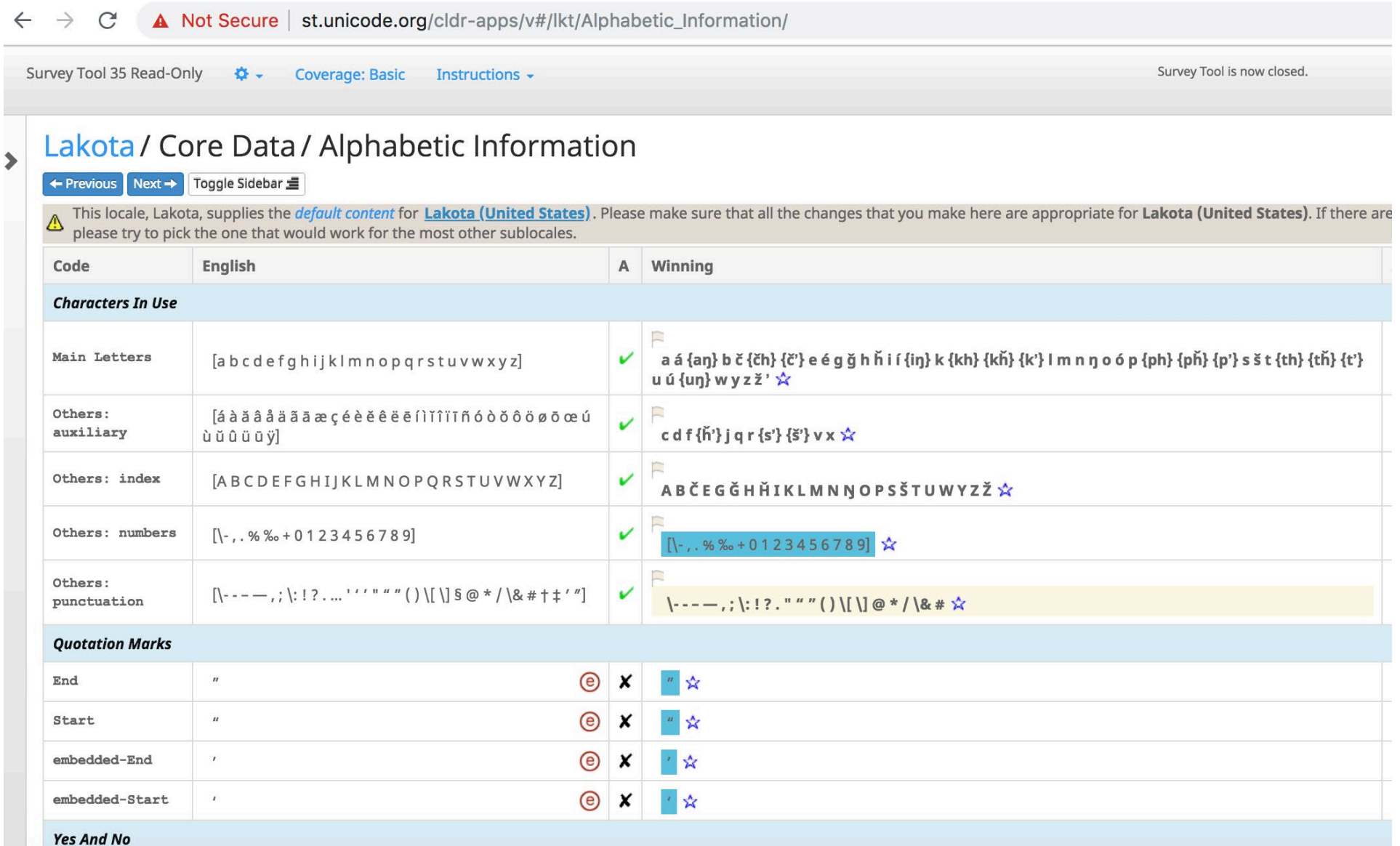
This data is to be entered using the Survey Tool except as noted.

1. (44+) 4 main Date/Time formats, 12 long&abbreviated, format&stand-alone month-names, 7 long&abbreviated day-names, 2 long day periods.
2. (01) Name of the language in the language.
3. (N) For any country locales, name of the country in the language, name/symbol for that country's currency. Must be at least one, for the default content locale.
4. (02) Datetime pattern, intervalFormatFallback
5. (05) (for Latn) decimal and grouping separators; decimal, currency, percent formats
6. (N) Names of countries (territories) with that language as official.
7. (M) Names of exemplarCities in multizone countries with that language as official
8. (05) Timezone patterns [http://cldr.unicode.org/translation/timezones]
9. (02) localePattern/Separator [http://cldr.unicode.org/translation/localepattern]
10. (03) key names
11. (14) long/short unit names (time intervals)

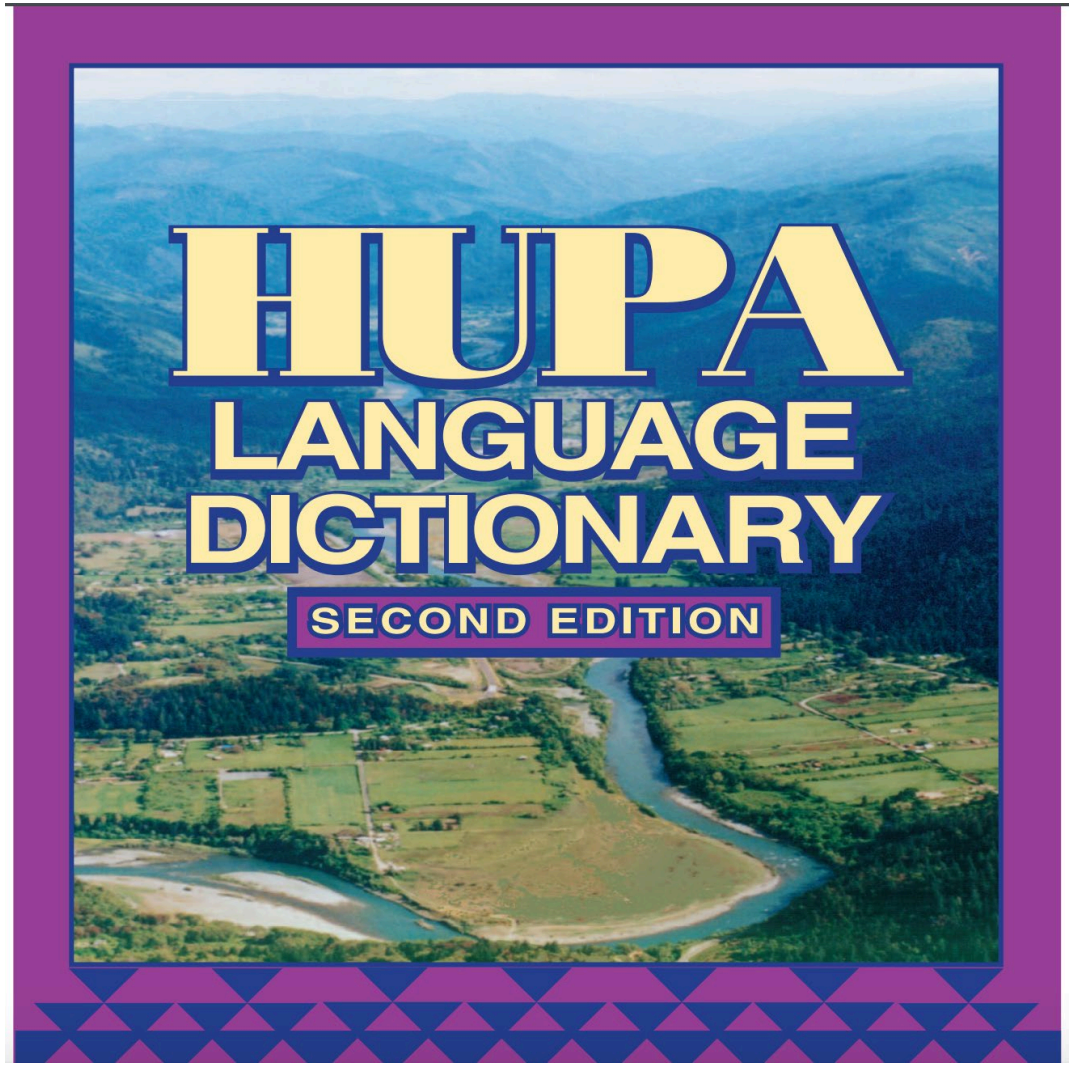
The First Seven Steps to ’Core XML Data’

1. Exemplar characters (dictionaries, grammar, references)
2. Left-to-right, Right-to-left, other directions
3. Plural rules (one duck, two ducks, etc.)
4. Script/region data (ISO 15924, 3166, 639)
5. More region data
6. Casing (titlecase for Latin-based languages)
7. Collation (fancy word for sorting; order with exemplars)

Getting To A Better CLDR Interface



Getting to Exemplars: Step 1A - Main Letters



HUPA ALPHABET CHART

a	a:	b	ch	ch’	chw
father	palm	bear	church	(**ch** with catch)	inchworm
whila’	whina:’	bo:se	mindich	which’ich’	chwich
(my hand)	(my eye)	(cat)	(bobcat)	(my elbow)	(firewood)

d	dz	e	e:	g	gy
deer	adze	met	men	geese	figure
dinday	didzit	whixe’	ne:s	niwhgit	digyun
(bullet)	(short)	(my foot)	(long)	(I’m afraid)	(here)

h	i	j	k	ky	k’
hen	hit	jar	keep	thank you	(**k** with catch)
xontah	mis	je:nis	king	kya’	k’ina’
(house)	(riverbank)	(day)	(stick)	(dress)	(Yurok)

ky’	l	ł	m	n	ng
(**ky** with catch)	let	(breathy **l**)	mill	now	ring
ky’oh	lah	ła’	milimil	nundil	whing
(porcupine)	(seaweed)	(one)	(flute)	(snow)	(song)

o	o:	q	q’	s	sh
tote	cone	(guttural **k**)	(**q** with catch)	sit	rush
dingq’och	to:-nehwa:n	qo	whiq’os	sa:ts’	nosht’ah
(sour)	(obsidian)	(worm)	(my throat)	(bear)	(I don’t believe it)

t	t’	tł’	ts	ts’	u
tea	(**t** with catch)	(**tł** with catch)	cats	(**ts** with catch)	run
to	t’e’	tł’oh	tse	ts’ił̇ting’	łixun
(water)	(blanket)	(grass)	(stone)	(weapon, rifle)	(sweet)

w	wh	x	xw	y	‘
word	whirred	(guttural **h**)	(guttural **wh**)	yes	(catch)
wiłdung’	wha	xong’	xwe:y	ya’	‘ah
(yesterday)	(sun)	(fire)	(his property)	(louse)	(cloud)

Getting to Exemplars: Step 1B - Index Letters

aachompa' store, town (1, 10)
aahámbi (irregular HNgr. of **aabi**) (8)
aahashtahli' window (10RS, 18)
aahilha' dance hall (18)
aahobachi to copy (something) from (a thing) (A,Noun,Noun) (16)
aaholloppi' graveyard, cemetery (18)
aaholhponi' pot (9RS), kitchen (10)
aahonkopa to steal (something) from (an institution) (A,Noun,Noun) (16)
aahopooni to cook (something) in, cook (something) on (A,Noun,Noun) (9RS)
Aaiklanna' part of **Nittak Hollo' Aaiklanna'**
aaiksaa to be made from (A,Noun) (13RS)
aaimpa' restaurant (10)
aaimpa' abooha dining room (10)
aaipita' in **issoba aaipita' plini'**
aaissachi to leave (A,S) (14RS)
aaittanaa' church (7RS, 10)
aaittapiha' group (20RS)
aaj'pa' table (10)
aaj'pa' holiito'pa' holy table (12RS)
aalhakoffichi to save from (A,S)
aalhpi'sa to be right, be correct; to do the right thing (A) (6)
aalhponi' kitchen (10)
aaminti to come from (A,Noun)
aanosi' bedroom (10)
aaombiniili' chair (12)
aapihlichika' kingdom (DP) (15RS)
aaponta to borrow (something) from (an institution) (A,Noun,Noun) (16)
aayimmi to believe in (something) (S;Noun) (17RS)

a

áa to go along, to be going along (A) (1, 20)
alhínchi to be enacted, made true (Noun) (17RS)
ashaka', ashka' behind, rear, back (SP); in back of (loc. n.) (SP) (10)
áshwa to stay, be there (A); to be located in (a place) (A,Noun) (dl. subj.) (12)
imáshwa to have (relatives) (dl. human obj.)

b

bala' bean (2)
sinti' bala' / bala' sinti' snake bean
balaafokha' / balaafka' pants (14)
banna to want (S,Noun) (9)
basha to be cut, to be operated on (S) (6)
itti' basha' abooha frame house
bashafa to be cut (of a long subj.) (S) (16)
bashaffi to cut (a long obj., usually with a knife) (A,S) (5)
bashli to operate on, to cut (A,S) (5)
bashpo knife (1, 7)
baafa to stab (A,S) (9)
bila to melt, to be melted (S) (1, 4)
biliili to melt (something) (A,S) (16)
billi'yacha billi'ya forever and ever (15RS)
bini'cha loshka sit down and tell lies! (an invitation to gossip) (3)
biniili to sit down, get into a sitting position (sg. subj.) (A) (12RS, 19)
bínni'li to sit, be seated, be in a sitting position (A); to sit in, to be located (sitting) in (a place) (A,Noun) (sg. subj.) (8, 10)
binohli to sit down, get into a sitting position (pl. subj.) (A) (19)
binohmáa to sit, be seated, be in a sitting position (A); to sit in, be located (sitting) in (a place) (A,Noun) (tpl. subj.) (10)
bíyyi'ka all the time, always (aux.) (1, 8); (aux. used with **±a'hi** "can" verbs) (5); it's all, it was all (7RS); truly (17RS); several, many (used as a stative verb following a noun, with focus endings) (18)
bo'li to hit (more than once); to beat up; to pound; to hit (pl. subj.) (A,S) (1, 13)
bo'wa to be hammered (Noun) (18)
bohli to lay down, put down (sg. long obj.) (A,S) (1, 13)
ishtombohli to accuse, blame; to accuse about, blame for [isht ombohli]
ombohli to lay (something) down on, put (something) down on

ch

chaffa one (3RS); to be one in number (A) (12); part of
nittak chaffaka one day

The Still Easy ‘Minimal’ Remainder

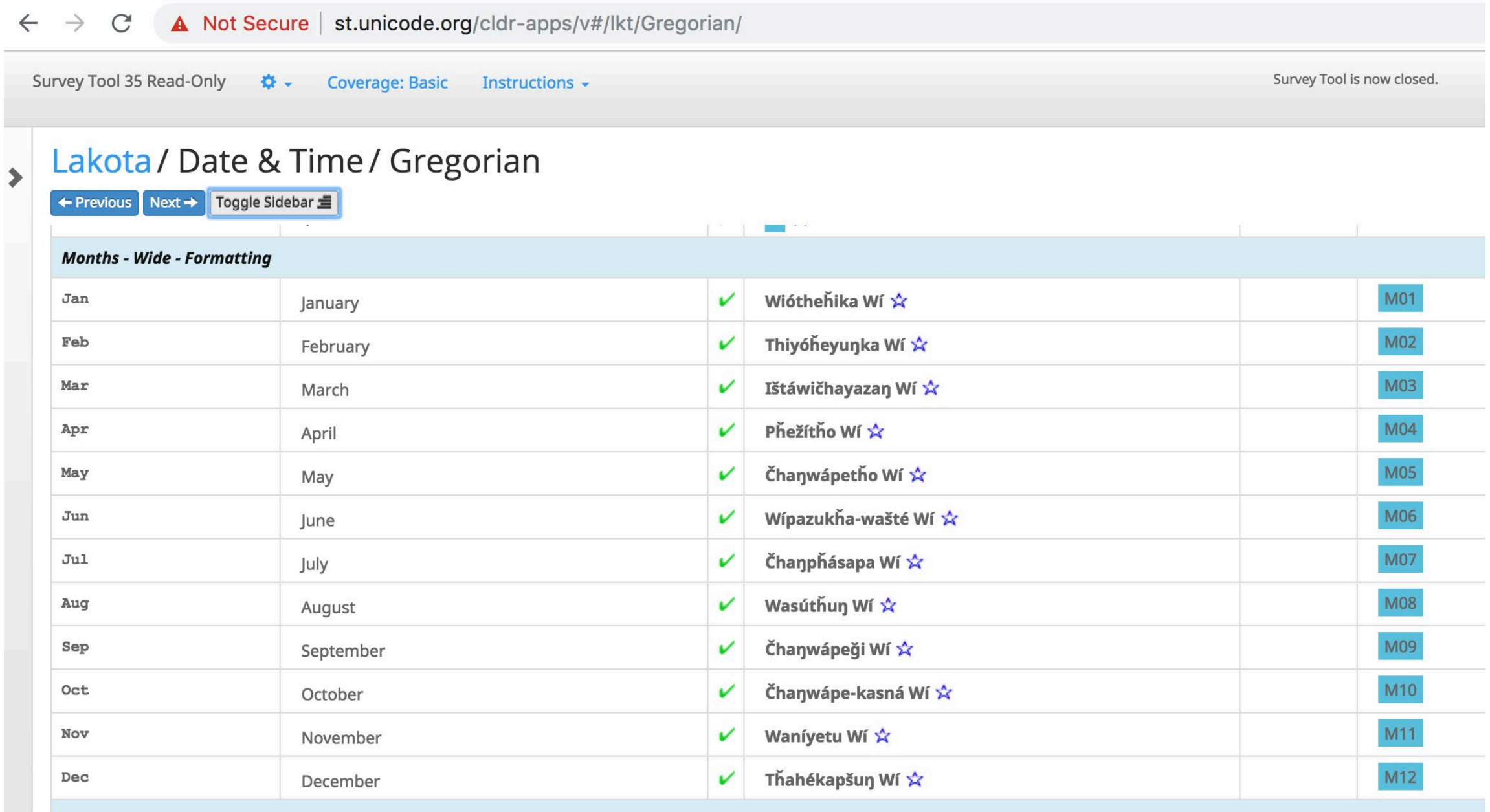
- Just Over 50 additional easy-to-find translations
  - (38) The brunt are some 12 month names and 7 day names
    - Translations often in dictionaries, apps, or other good references
  - (1) Name of the language in that language
  - (7) “Year”, “Month”, “Week”, “Day”, “Hour”, “Minute”, “Second”
  - (6) Next slide

- With caveats
  - Some translations may require a bit more ‘wordsmithing’
  - Abbreviations may not be a concept in a given language

The Last Six Strings To Minimal

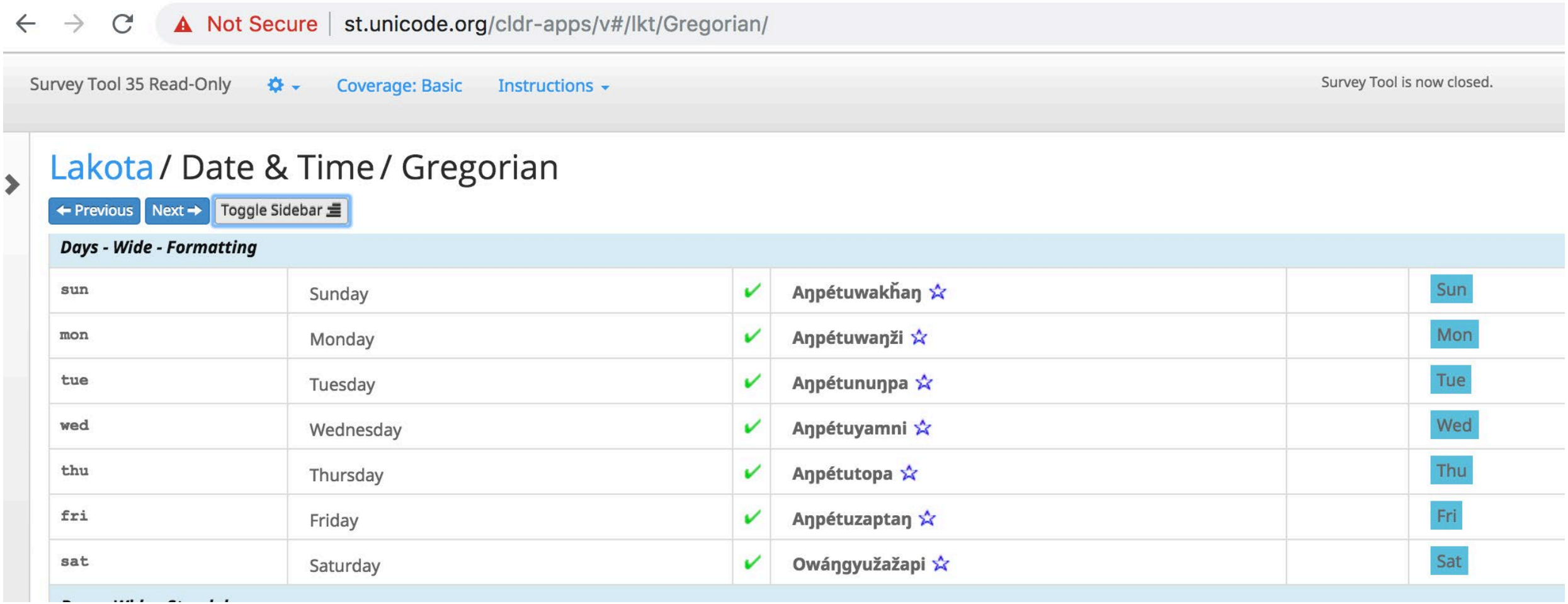
- “Standard Time”
- “Daylight Time”
- “Gregorian Calendar”
- “Standard Sort Order”
- “Western Digits”
- <date> “at” <time>

Months of the Year

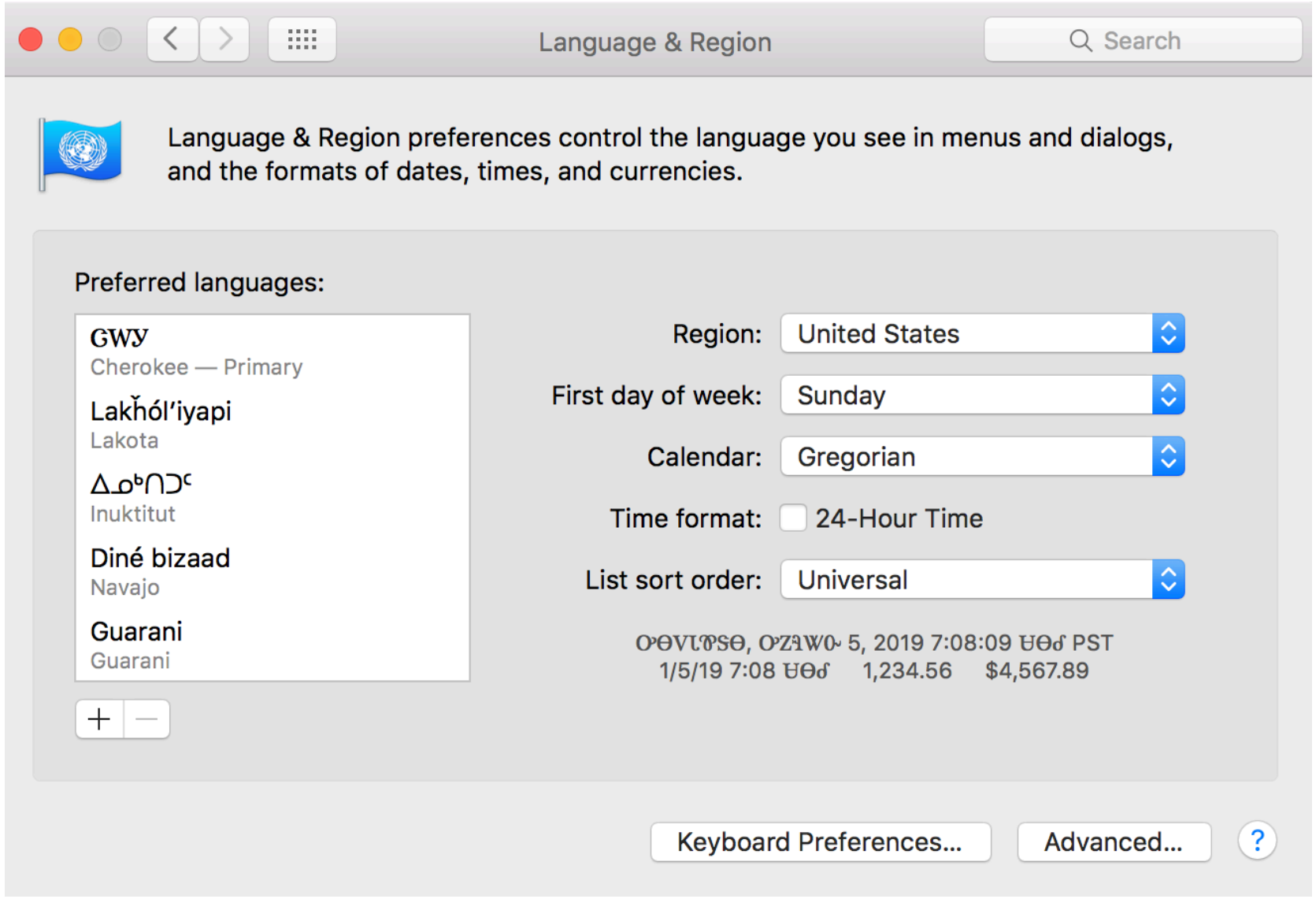


Months - Wide - Formatting					
Jan	January	✔	Wióthehika Wí ☆		M01
Feb	February	✔	Thiyóȟeyuŋka Wí ☆		M02
Mar	March	✔	Ištáwičhayazaŋ Wí ☆		M03
Apr	April	✔	Pȟežítȟo Wí ☆		M04
May	May	✔	Čhaŋwápetȟo Wí ☆		M05
Jun	June	✔	Wípazukȟa-waště Wí ☆		M06
Jul	July	✔	Čhaŋpȟásapa Wí ☆		M07
Aug	August	✔	Wasútȟuŋ Wí ☆		M08
Sep	September	✔	Čhaŋwápeği Wí ☆		M09
Oct	October	✔	Čhaŋwápe-kasná Wí ☆		M10
Nov	November	✔	Waníyetu Wí ☆		M11
Dec	December	✔	Tȟahékapšuŋ Wí ☆		M12

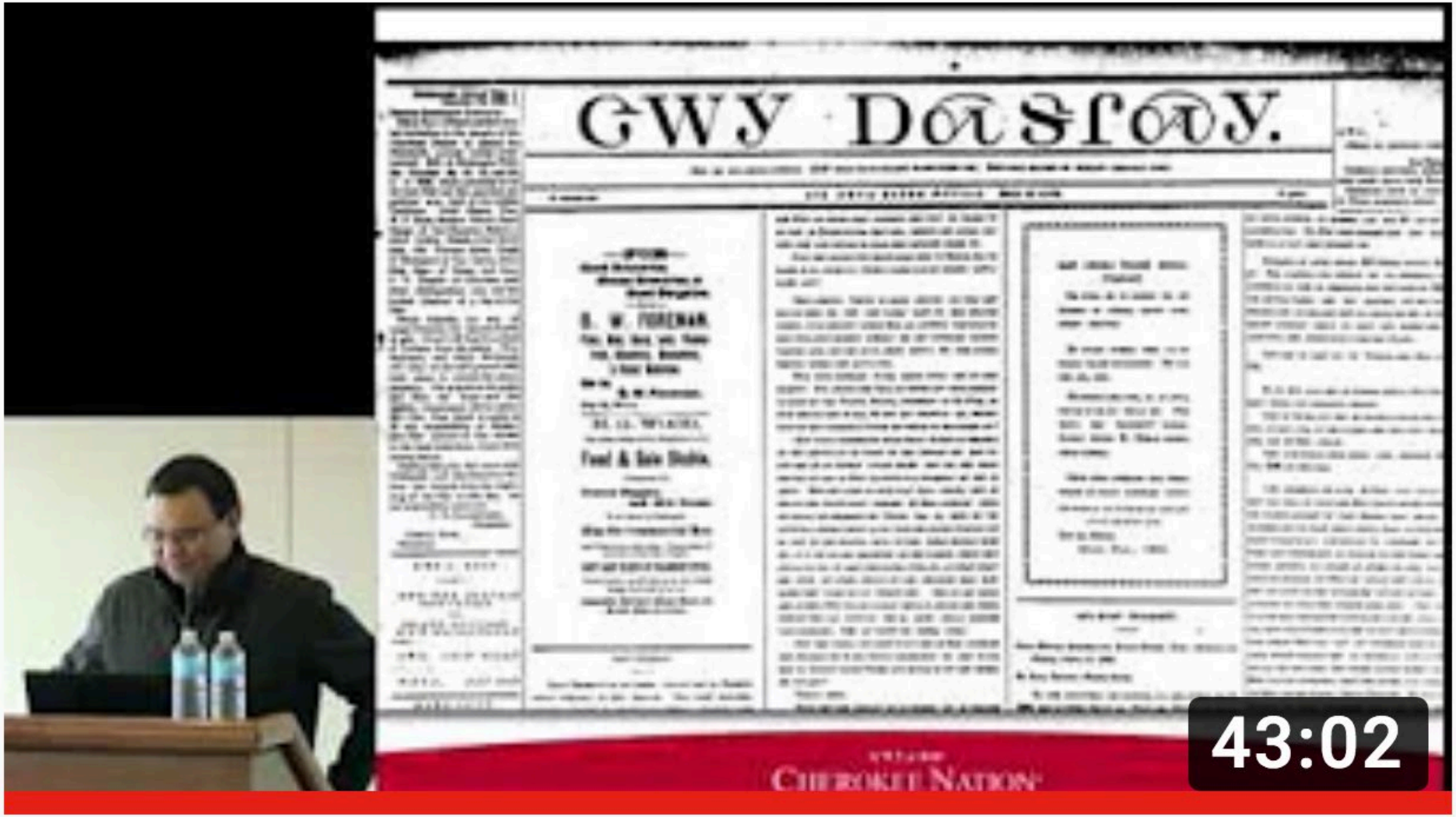
Days of the Week



What Success Looks Like On A Device



A Success Story: Cherokee (in their own words)



https://youtu.be/EEEu8ufwW08?t=1498

How To Get Started

- File an new ticket with the ‘Core XML Data’ here:
  - http://unicode.org/cldr/trac
  - For reference, Osage, Muscogee/Creek, Chickasaw examples here:
    - https://unicode.org/cldr/trac/ticket/10721
    - https://unicode.org/cldr/trac/ticket/11424
    - https://unicode.org/cldr/trac/ticket/11983
- Ping me with ticket #
  - craig@unicode.org

[PJC] Lecture 22 Saving the World with Computing

Q&A

craig@unicode.org

ccummings@vmware.com

Postscript

- CLDR Release Timing
  - Spring and fall of each year; fall is a bigger release
  - Calendar on CLDR home page
- Unencoded scripts require more work (e.g., Mi’kmaq)
  - Happy to guide towards a Unicode character encoding proposal
- Patience is required
  - Data may only take an hour to prepare, but keyboards, fonts, and system integrations will take longer.
- Full localizations of user interfaces, online help, and other documentation are a different discussion
  - Happy to have that discussion, too.