



Innovations in i18n at Google

Mark Davis
Vladimir Weinstein

Making the Multilingual Web Work, W3C Workshop
12 March 2013

Core Internationalization

- Unicode [Encoding](#)
- Unicode [CLDR](#)
- [ICU](#)

Beyond the core

- Segmentation, Character Encoding Detection,...
- Entities, Names, Phone Numbers, Addresses
- Plurals/Gender, Languages, Translate, Speech
- Fonts, Localization, Input

Entity i18n

Entities now being used in Google Search (e.g. Knowledge Panel)

...but what about other languages?

(en)Computer Science

(fr) Informatique

(de)Informatik

(it) Informatica

(sr) Информатика

(he)מדעי המחשב

(zh) 计算机科学

...



- Entities are locale-independent
- Entity names are locale- / context-specific
- Links are locale-independent, but the strength of links is locale-specific
- Categories tend to be locale-specific (e.g. categorization of colors)

Names

- Personal Names: Barack “Barry” Obama
- Google+ Pages: “[Lindt Chocolate World](#)”
- Custom URLs: google.com/+toyota

Names: Unicode Security Issues

- No* mixed scripts: [paypal](#) problem
- No mixed numbering systems
 - U+09EA (8) BENGALI DIGIT FOUR
 - U+0038 (8) DIGIT EIGHT
- [Too many accents](#), duplicates
 - a^ˆ, a^ˆ
 - Maaaark

See [UTR #39: Unicode Security Mechanisms](#), [ICU](#)

Names: Normalization

- Display vs Identity
- Google+ Pages: more freeform
- Unicode compatibility mappings
 - $o'' \rightarrow \ddot{o}$ (or o for Custom URLs)
 - $AA\square\square\square \rightarrow A$
- Remove invisible characters; replace whitespace
- URLs: Replace non-letters/marks by '-'; drop accents; uniqueness
- ...

Names: Formatting

English

{GIVEN}
 {GIVEN} "{NICK}"
 {GIVEN} ({NICK})

{GIVEN} {FAMILY}
 {GIVEN} "{NICK}" {FAMILY}
 {GIVEN} {FAMILY} ({NICK})

{FAMILY}, {GIVEN}
 {FAMILY}, {GIVEN} "{NICK}"
 {FAMILY}, {GIVEN} ({NICK})

Japanese

{GIVEN}
 {GIVEN}{SP}"{NICK}"
 {GIVEN}{SP}({NICK})

{GIVEN}{SP}{FAMILY}
 {GIVEN}{SP}"{NICK}"{SP}{FAMILY}
 {GIVEN}{SP}{FAMILY}{SP}({NICK})

{FAMILY}{SP}{GIVEN}
 {FAMILY}{SP}"{NICK}"{SP}{GIVEN}
 {FAMILY}{SP}{GIVEN}{SP}({NICK})

Plurals & Gender

Examples:

- The Agent: Alice added 1 people to his circle.
- The User: Vous êtes le seul participant à l'appel.

Hurdles

- Different messages to different translators
- Most translators are not software engineers
- Most engineers don't speak 60 languages
- May not know the gender
- Languages vary:
 - English (2), French (2*), Russian (4), Arabic (6),...

Plurals & Gender: progress

Plurals — with numbers written as digits

- Cardinals (1, 2,...); Ordinals (1st, 2nd,...)
- Messages
- Currencies, units, compact numbers

Genders — for (known) people, lists of people

1 milijon
2 milijona
3 milijone
5 milijonov
101 milijon

Lists

- Switzerland **and** Japan
- Switzerland, Japan, Egypt, **and** Canada

Message 1

Message ID:

Message text

[PLURAL_TOTAL_USERS]

0 (nombre exact)

[=0]No one liked this.

1 (nombre exact)

[=1]One person liked this.

0, 0.5, 1, ...

[ONE]# people liked this.

2, 2.5, 3, ...

[OTHER]# people liked this.

[END_PLURAL]

Description

[ICU Syntax] Who liked this? This message requires special attention. Please follow the instructions here:

Placeholders



Message 1

Message ID:

Message text

[PLURAL_TOTAL_USERS]

0 (nombre exact)

{0/} personne n'a aimé cela

Characters: 23[Next»](#)

Actions ▾

1 (nombre exact)

[=1]One person liked this.

0, 0.5, 1, ...

[ONE]# people liked this.

2, 2.5, 3, ...

[OTHER]# people liked this.

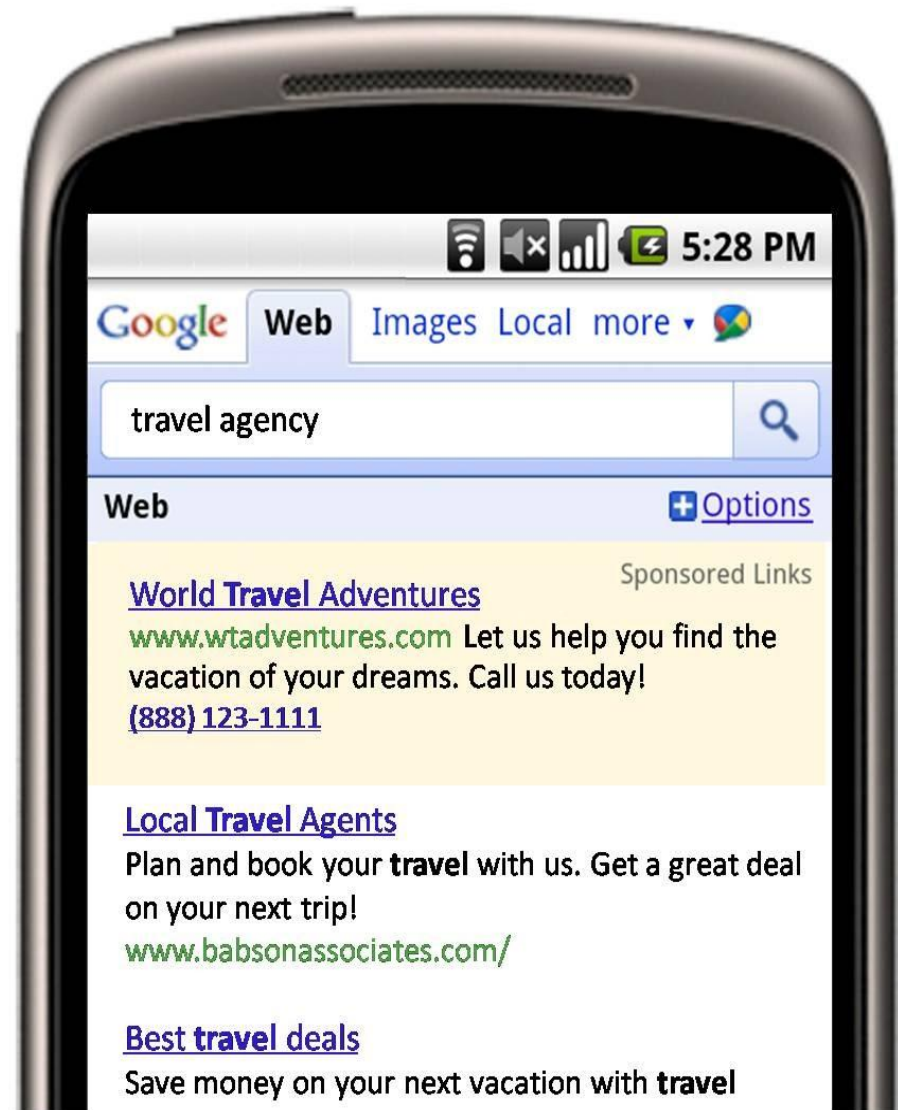
[END_PLURAL]

Phone Numbers

(011) 1234 5678

(011) 15 1234 5678

+54 9 11 1234 5678



Phone Number Library

<http://code.google.com/p/libphonenumber/>

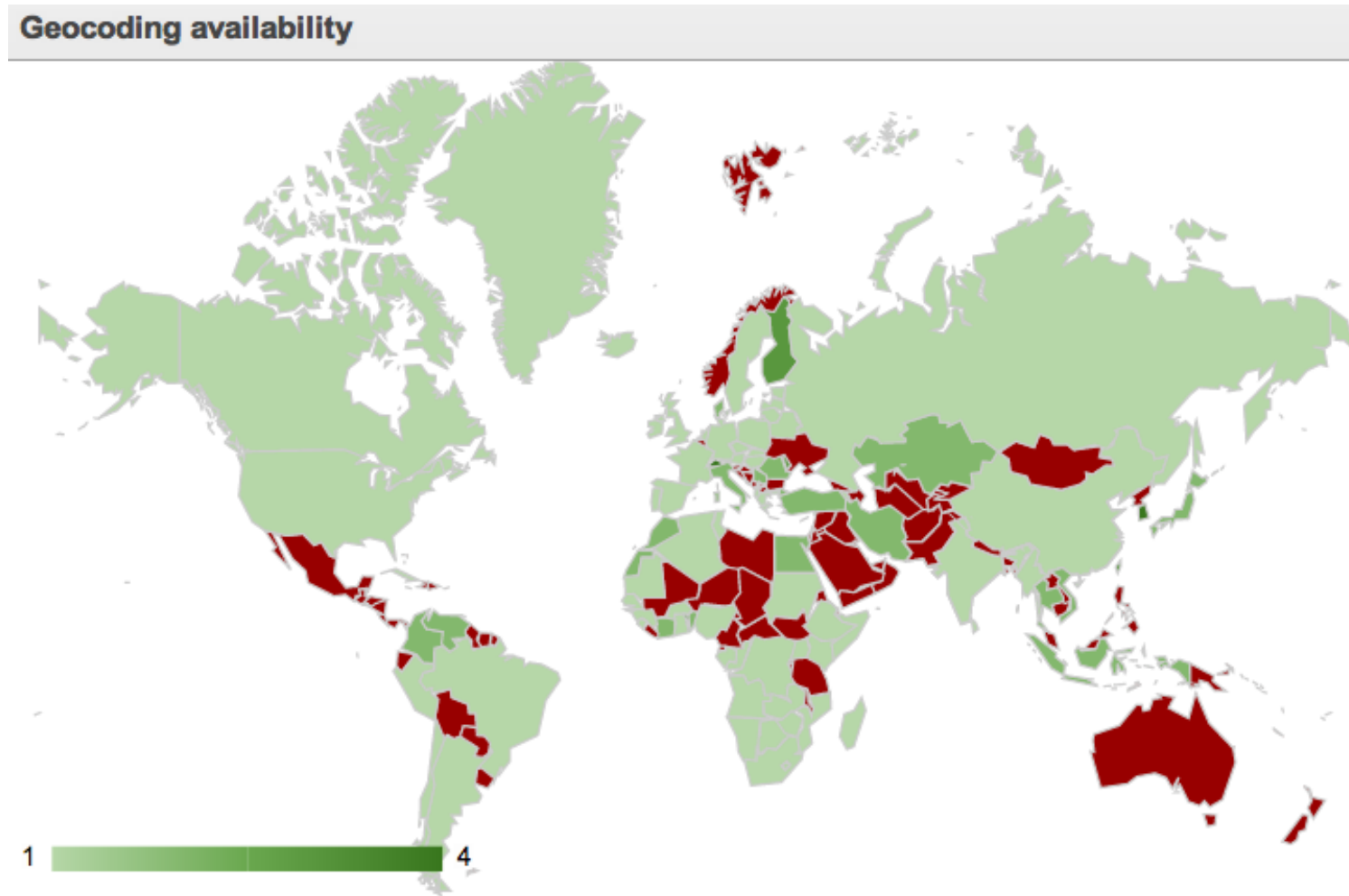
Handles...

- parsing/formatting/validating/matching/canonicalization
- getting types/example numbers
- finding numbers in text

Users

- Google: Android, Google Voice, Chrome, AdWords, Google Maps, Gmail Contacts, etc.
- Others: Facebook Messenger, Tango, RH Brands, Guardly, Thrutu, etc.
- Open source

Geocoding in Android (ICS+)



Addresses

<http://code.google.com/p/libaddressinput/>

Handles...

- detailed validation for many regions
- layout and basic validation for all regions

Users

- Google: Android, Google Checkout, Maps,...

A man in a brown coat and scarf is speaking into a microphone with a logo. He is surrounded by a crowd of people, some wearing dark clothing and hats. The background shows a brick building and a banner.

Google

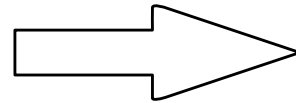
got my language settings wrong

Bad Language User Experience

- Pick language...other products still default to IP, etc



English speaker in Germany signs up for Google+ in English



Google Calendar shows up in German

- Especially when desired language \neq default for IP:
 - Travelling
 - Don't speak majority language (immigrants/expats)
 - Multiple-language regions (Switzerland, India...)
 - VPN/Redirect

Worse for Enterprise

Google Apps for Your Domain Admin sets default language in Control Panel...



(Until recently)
Only Gmail works!

Universal Language Settings

Google Properties Using ULS



Google+



Gmail



Apps Control Panel



Accounts



Groups



Play



Translate



Search



Drive/Docs



Maps

www.google.com/settings/language

Why >1 Language?

- **Fallback:** User's top choice may not be available in all products
- **Content:** For *all* languages a user speaks, we can show posts, search results, autotranslate into

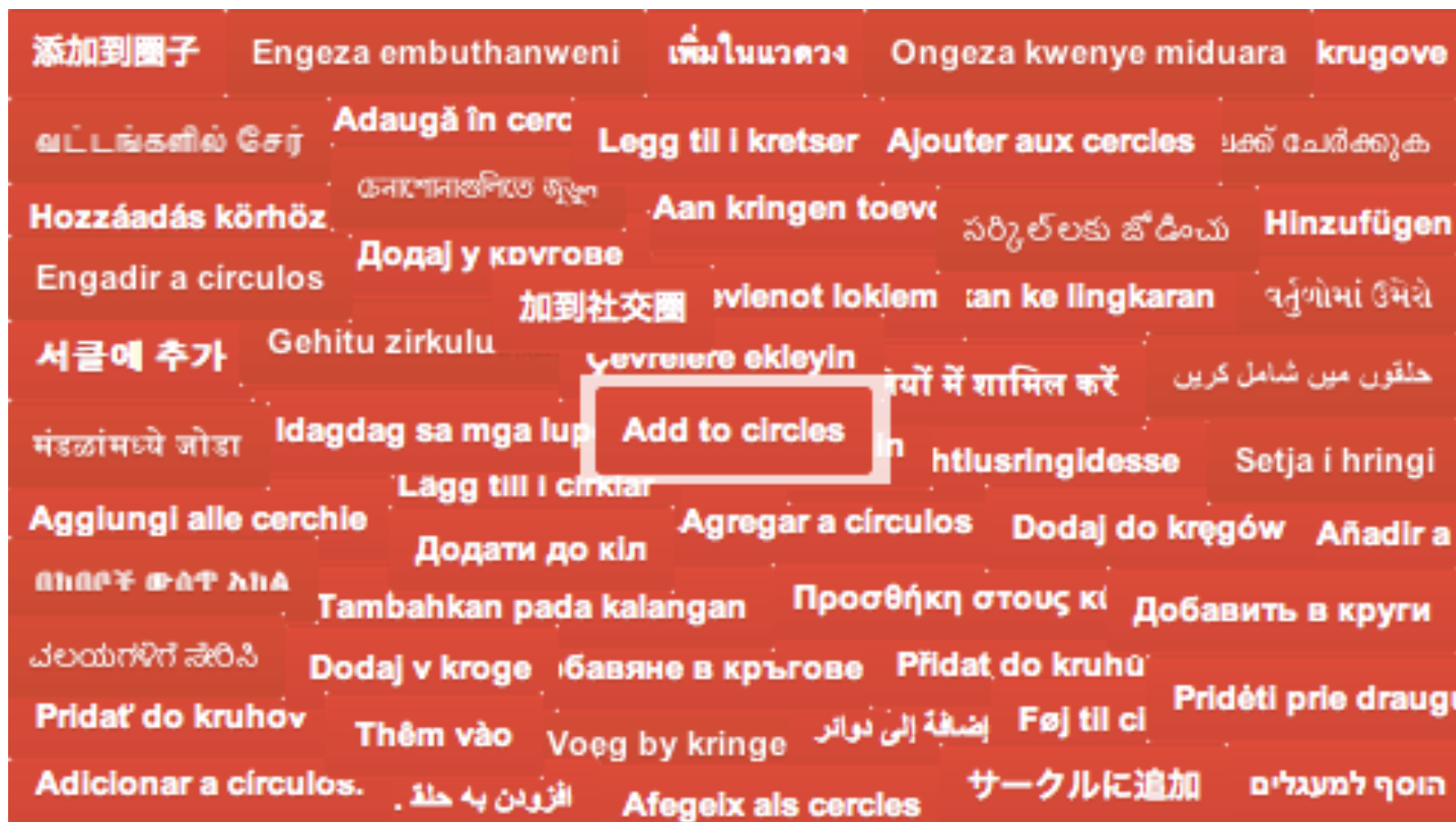
Default languages

Specify languages in order of preference for viewing Google products.

1. English (United States) - English (United States) ▼ Primary language
2. French (France) - français (France) ▼ [Make primary](#) · [Remove](#)
3. Arabic - العربية ▼ [Make primary](#) · [Remove](#)
4. Italian - italiano ▼ [Make primary](#) · [Remove](#)

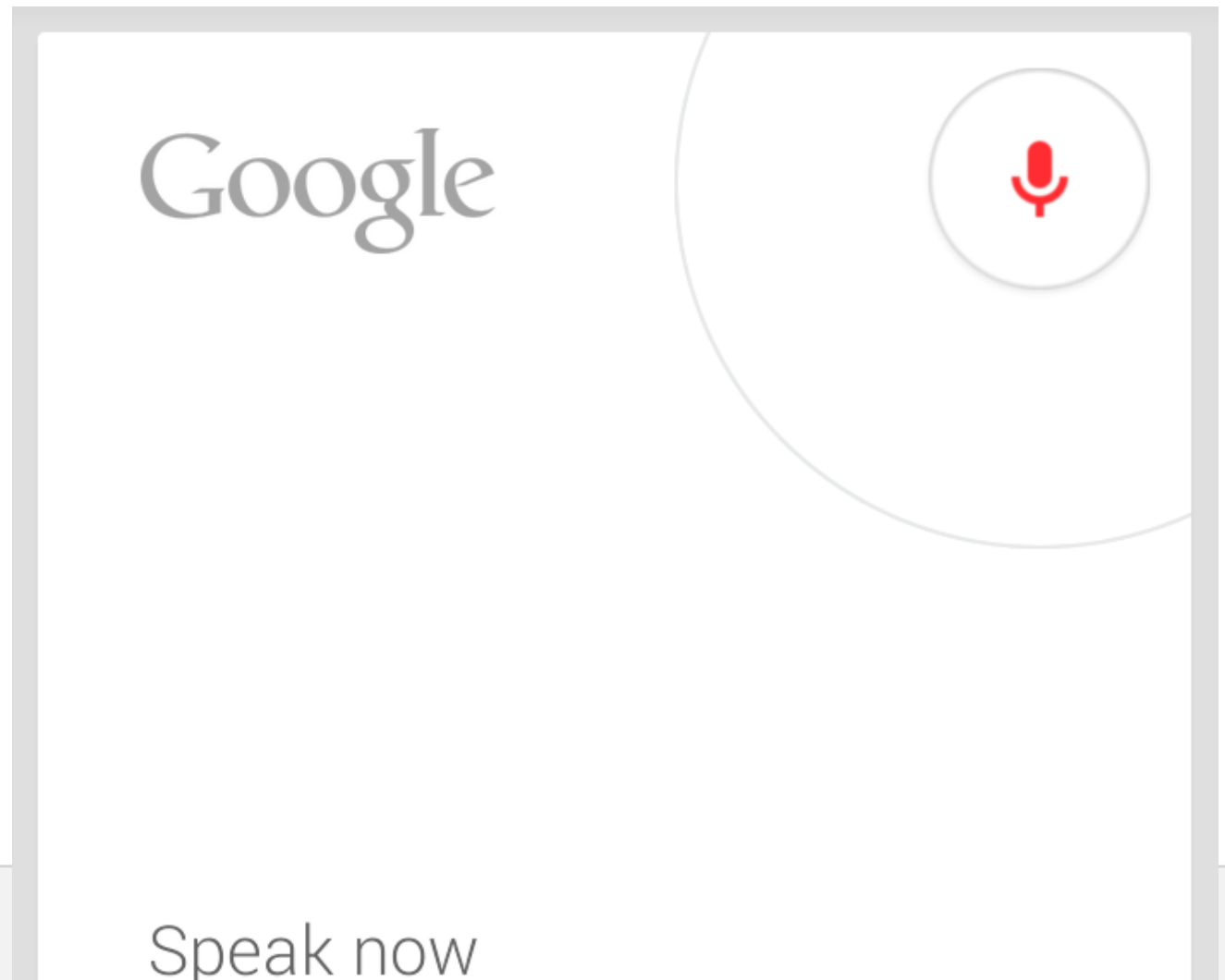
[+ Add another language](#)

60 Language Initiative



Google Speech-to-Text

Now supports 42 languages, including accents/dialects in 46 countries



Input Tools

<http://www.google.com/inputtools/>



Translate

From: Bengali



To: French

Translate

English Bengali French Detect language

বিড়াল



বা



French

chat



namaste

1. नमस्ते
2. नमसते
3. नमास्ते
4. नामास्ते
5. नामस्ते
6. namaste



Google Search

Font Development at Google

We want to remove tofu on the Web!

neñu • Fiji Hindi • Føroyskt • Fransch • Furlan • Gaelg • Gagauz • Gàidhlig • 贛語 • گیلکی
ar • Kalaallisut • □□□□□□□□□□ • Kaszëbsczi • Kernewek • □□□□□□ • Kinyarwanda • K
in • Malti • 文言 • Māori • □□□□□□□□□□ • □□□□ • □□□□□□□□□□□□ □ • Мокшень • Монгол
• □□□□□□ • पाळि • Pangasinán • پنجابى / پښتو • Papiamentu • پښتو • Перем Коми • Pfälzi
язык • □□□□□□ • Sámeigiella • Sardu • Саха Тыла • Scots • Seeltersk • සිංහල • Ślůnski •
г • Удмурт • Uyghur / □□□□□□ • Vèneto • Vöro • West-Vlams • Wolof • 吴语 • ייִדיש • Zeê

Pan-Unicode Font (Noto = No Tofu)

<http://code.google.com/p/noto/>

Sfntly Font Library

<http://code.google.com/p/sfntly/>

Supports

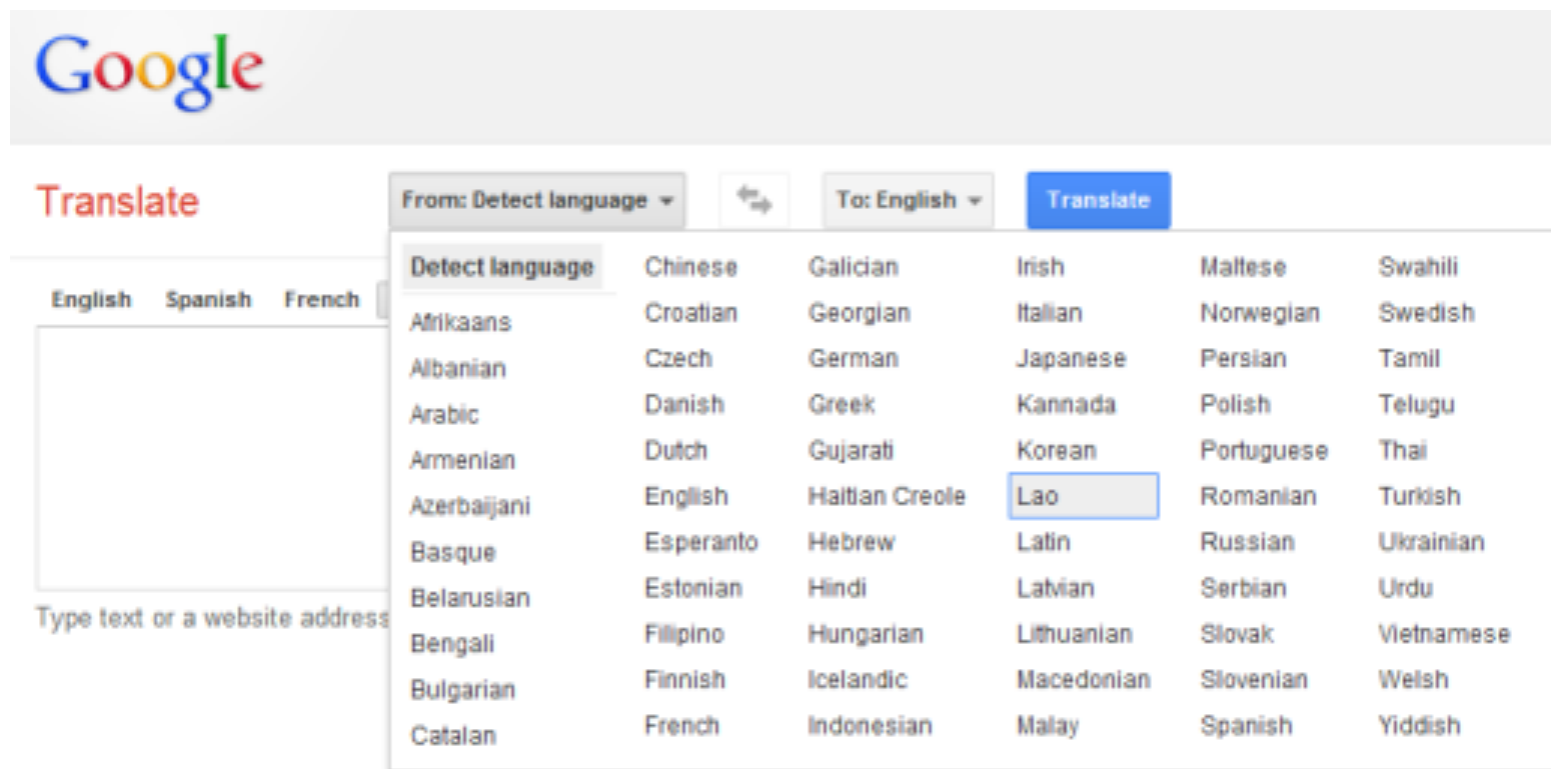
- reading and writing basic font tables
- support for bitmap glyphs
- e.g. can use to subset fonts, find problems, extract information

Used

- Google Web Fonts: manipulating/subsetting fonts
- Chrome/Google Cloud Print: used to subset fonts for PDF printing
- Google+: measuring text length
- ...and various other Google/non-Google tools

Google Translate

Now supports 65+ languages



Google Localization Infrastructure


<http://translate.google.com/toolkit>

...now used for *all* of Google's localization!

Original text:
Translation: English to Chinese (Simplified Han)
2% complete, 3189 w

Great Pyramid of Giza


[0]



The Great Pyramid of Giza (also called the Khufu's Pyramid, Pyramid of Khufu, and Pyramid of Cheops) is the oldest and largest of the three pyramids in the Giza Necropolis bordering what is now Cairo, Egypt, and is the only one of the Seven Wonders of the Ancient World that survives substantially intact. It is believed the pyramid was built as a tomb for Fourth dynasty Egyptian King Khufu (Cheops in Greek) and constructed over a 20 year period concluding around 2560 BC. The Great Pyramid was the tallest man-made structure in the world for over 3,800 years. Originally the Great Pyramid was covered by casing stones that formed a smooth outer surface, and what is seen today is the underlying core structure. Some of the casing stones that once covered the structure can still be seen around the base. There have been varying scientific and alternative theories regarding the Great Pyramid's construction techniques. Most accepted construction theories are based on the idea that it was built by moving huge stones from a quarry and dragging and lifting them into place.

There are three known chambers inside the Great Pyramid. The lowest chamber is cut into the bedrock upon which the pyramid was built and was unfinished. The so-called Queen's Chamber and King's Chamber are higher up within the pyramid structure. The Great Pyramid of Giza is the main part of a complex setting of buildings that included two mortuary temples in honor of Khufu (one close to the pyramid and one near the Nile), three smaller pyramids for Khufu's wives, an even smaller "satellite" pyramid, a raised causeway connecting the two temples, and small mastaba tombs surrounding the pyramid for nobles.

[0]



伟大的埃及金字塔（也称胡夫金字塔，胡夫金字塔，和金字塔胡夫）是历史最悠久，规模最大的三个在吉萨金字塔群中的建筑，埃及开罗，并是唯一的一个七世纪的古代世界。金字塔大大完好无损。据认为建造金字塔的吉萨为第四王朝埃及国王胡夫（希腊语称胡夫）所建造的20年内的2560年的结论公元前。大金字塔是最高的建筑体在世界上超过三千八百年来。原来是大金字塔所建造者胡夫形成并表面光滑，并我们今天所看到的是基本准的核心结构。一些都管石头。一旦涉及到结构仍然可以看到周围的基础。有不同的科学和替代理论关于大金字塔的建造技术。大多数视觉理论是基于一种观点，称它是由移动巨大的石块从采石场和拖拽与翻转到位。

有三个已知部分内的大金字塔，最低分插入基岩胡夫金字塔的建造工作。完成，所谓的“皇后”室和两个并排并排胡夫（一靠近金字塔和一个附近的尼罗河），三个小的胡夫金字塔的孪子，一个更小的“卫星”金字塔，一个突出去或连接两个寺庙，和小mastaba金字塔坟墓周围的贵族。

建筑金字塔

ARB Format for Web App Localization

Application Resource Bundles

<http://code.google.com/p/arb/wiki/ApplicationResourceBundleSpecification>

```
arb.register("arb_ref_app", {
  "@@locale": "en_US",
  "@@context": "HomePage",

  "MSG_OK": "Everything works fine.",           // simple message

  "FOO_123": "Your pending cost is {COST}", // with placeholder
  "@FOO_123": {                               // and annotations
    "type": "text",
    "context": "HomePage:MainPanel",
    "description": "balance statement.",...
  }
}
```



© Disney/Pixar Animation Studios. All Rights Reserved

YouTube Caption Translation

Kenji chuting clothes Last edit was made 9 days ago by Unknown user

Show toolkit

Publish to YouTube

File Edit View Help



Translation: English » Chinese (Simplified)

33% complete, 45 words

00:00:03,610 - 00:00:10,610 Kenji likes chuting his own clothes

↗ ✕

健二喜欢自由落下自己的衣服

Characters: 13

▶
↺
🖌️
⌂
◀
▶

00:00:29,490 - 00:00:33,280 Just say "Chute clothes"
只是说“滑道衣服”

00:00:33,280 - 00:00:34,249 and he'll pick up his clothes
他会拿起他的衣服

00:00:34,249 - 00:00:41,249 and then run to the laundry chute
, 然后运行到洗衣槽

00:00:41,660 - 00:00:47,760 laughing and giggling all the way
一路笑, 傻笑

00:00:47,760 - 00:00:54,760 it's so cute!
它是如此的可爱!

And even more...

- *HarfBuzz*
- *ICU / CLDR*
- *JavaScript (Internationalizing the Core)*
- *Emoji*
- *Bidi and RTL*
- *I18n Testing*

Questions?

