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

Google

Names: Unicode Security Issues

- No* mixed scripts: paypal problem
- No mixed numbering systems
  - U+09EA (৪) BENGALI DIGIT FOUR
  - U+0038 (8) DIGIT EIGHT
- Too many accents, duplicates
  - a¨ˆ`, a¨ˆ`
  - Maaaark

See *UTR #39: Unicode Security Mechanisms*, ICU

Google

Names: Normalization

- Display vs Identity
- Google+ Pages: more freeform
- Unicode compatibility mappings
  - o¨ → ö (or o for Custom URLs)
  - AA□□□ → A
- Remove invisible characters; replace whitespace
- URLs: Replace non-letters/marks by '-'; drop accents; uniqueness
- …

Google

Names: Formatting

English	Japanese
{GIVEN}	{GIVEN}
{GIVEN} “{NICK}”	{GIVEN}{SP}“{NICK}”
{GIVEN} ({NICK})	{GIVEN}{SP}（{NICK}）
{GIVEN} {FAMILY}	{GIVEN}{SP}{FAMILY}
{GIVEN} “{NICK}” {FAMILY}	{GIVEN}{SP}“{NICK}”{SP}{FAMILY}
{GIVEN} {FAMILY} ({NICK})	{GIVEN}{SP}{FAMILY}{SP}（{NICK}）
{FAMILY}, {GIVEN}	{FAMILY}{SP}{GIVEN}
{FAMILY}, {GIVEN}“{NICK}”	{FAMILY}{SP}“{NICK}”{SP}{GIVEN}
{FAMILY}, {GIVEN}({NICK})	{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),…

Google

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

Lists

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

1 milijon
2 milijona
3 milijone
5 milijonov
101 milijon

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]

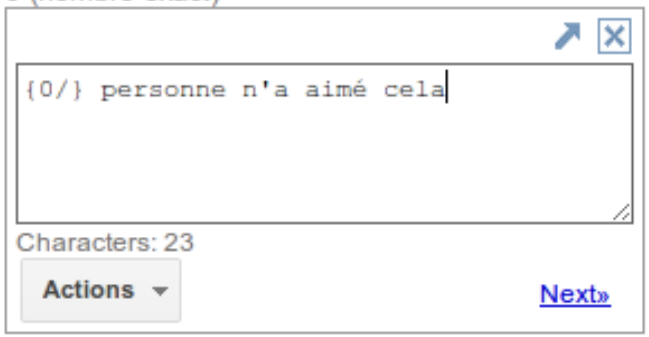
Description
[ICU Syntax] Who liked this? This message requires special attention. Please follow the instructions here:

Placeholders

Message 1 **Message ID:**

Message text

[PLURAL_TOTAL_USERS]
0 (nombre exact)



1 (nombre exact)
[=1]One person liked this.

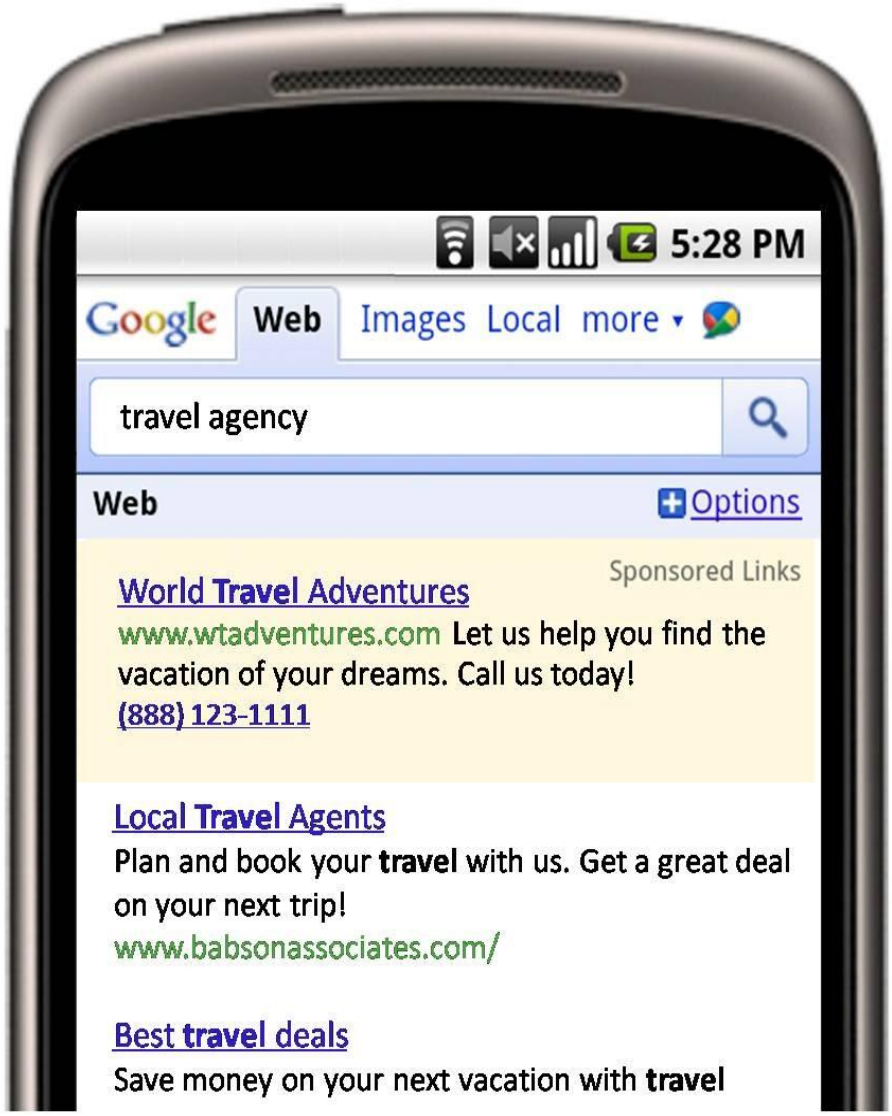
0, 0.5, 1, ...
[ONE]# people liked this.

2, 2.5, 3, ...
[OTHER]# people liked this.
[END_PLURAL]

Google

Phone Numbers

(011) 1234 5678
(011) 15 1234 5678
+54 9 11 1234 5678



Phone Number Library

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

Handles...

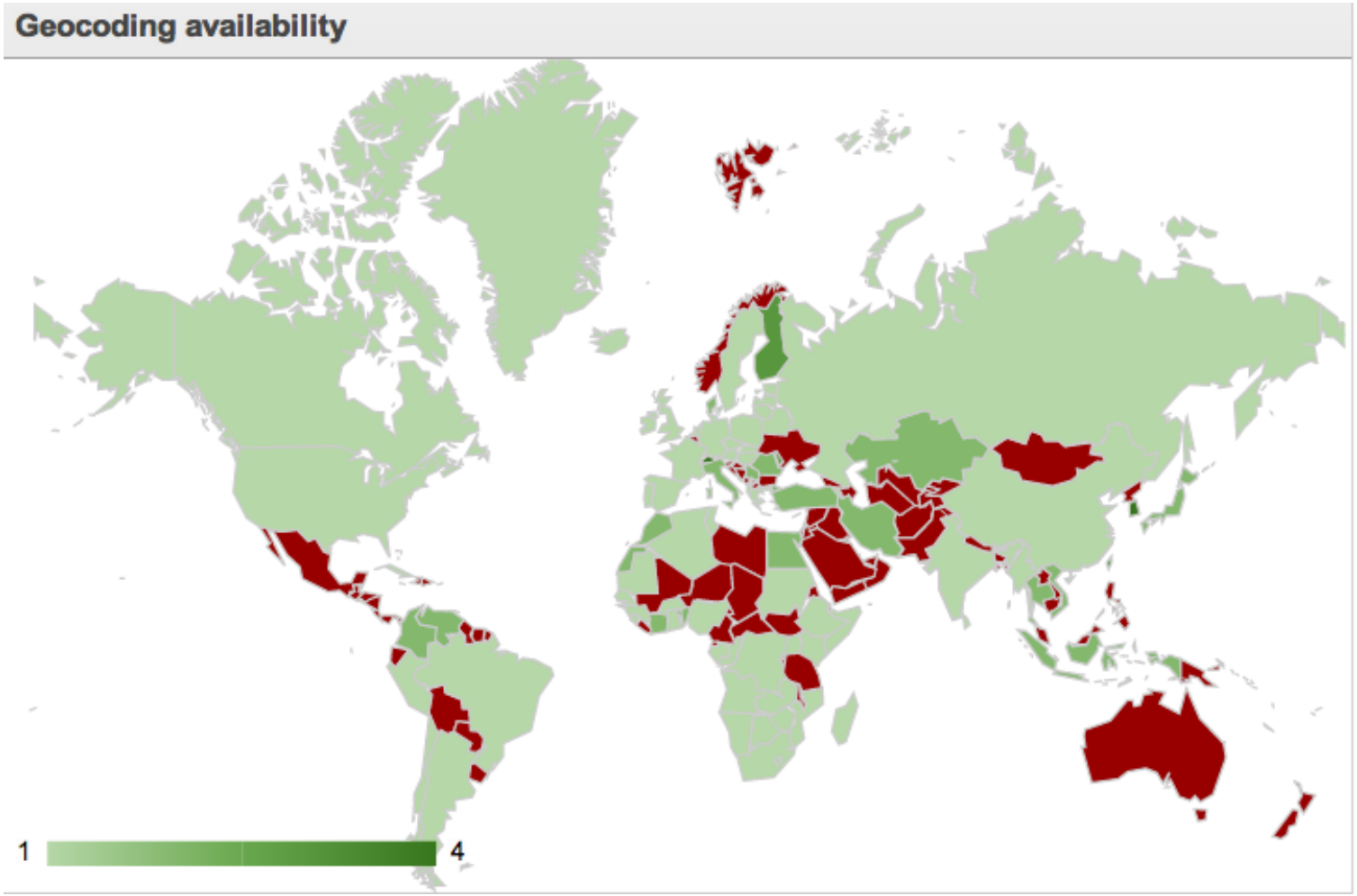
- parsing/formatting/validating/matching/canonicalization
- getting types/example numbers
- finding numbers in text

Users

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

Google

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

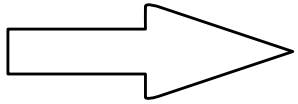
- Google: Android, Google Checkout, Maps,…


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 ≠ 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!

Google

Universal Language Settings

Google Properties Using ULS

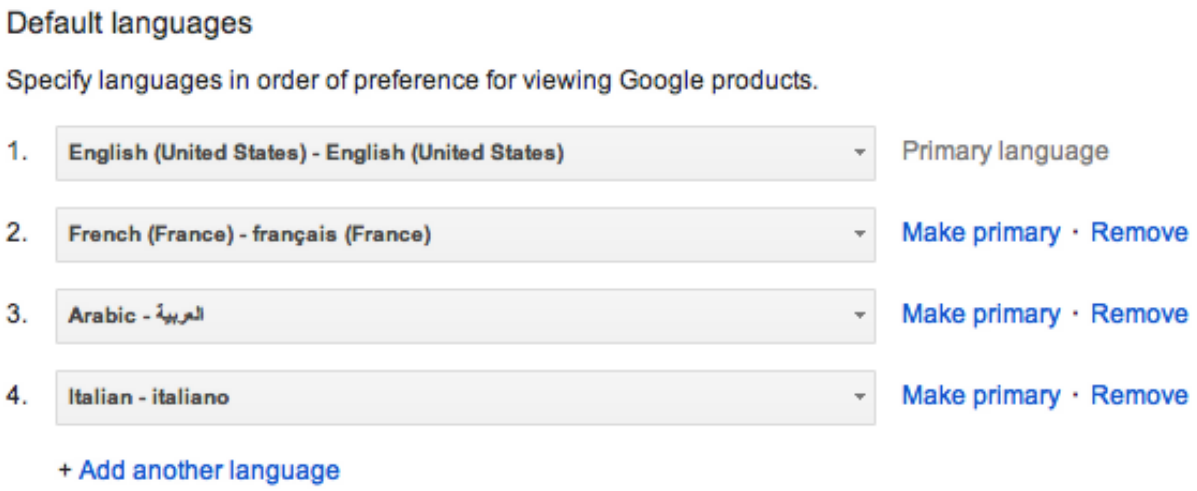
- Google+
- Gmail
- Apps Control Panel
- Accounts
- Groups
- Play
- Translate
- Search
- Drive/Docs
- Maps

www.google.com/settings/language

Google

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



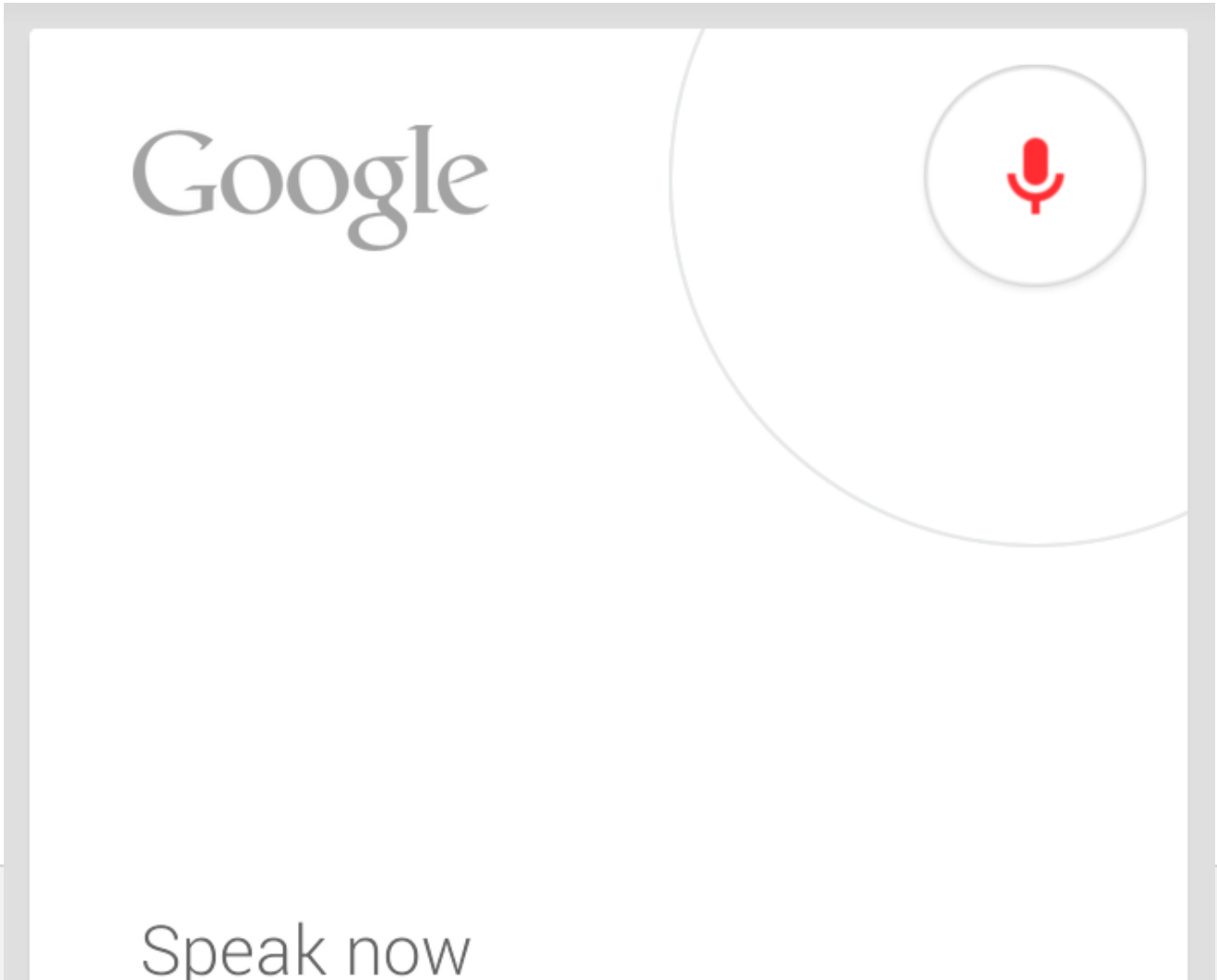
Google

60 Language Initiative

Google

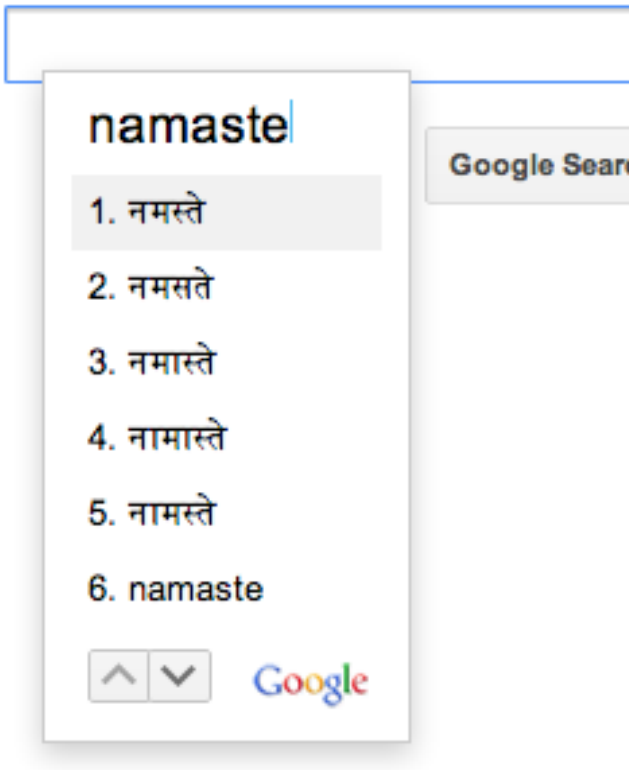
Google Speech-to-Text

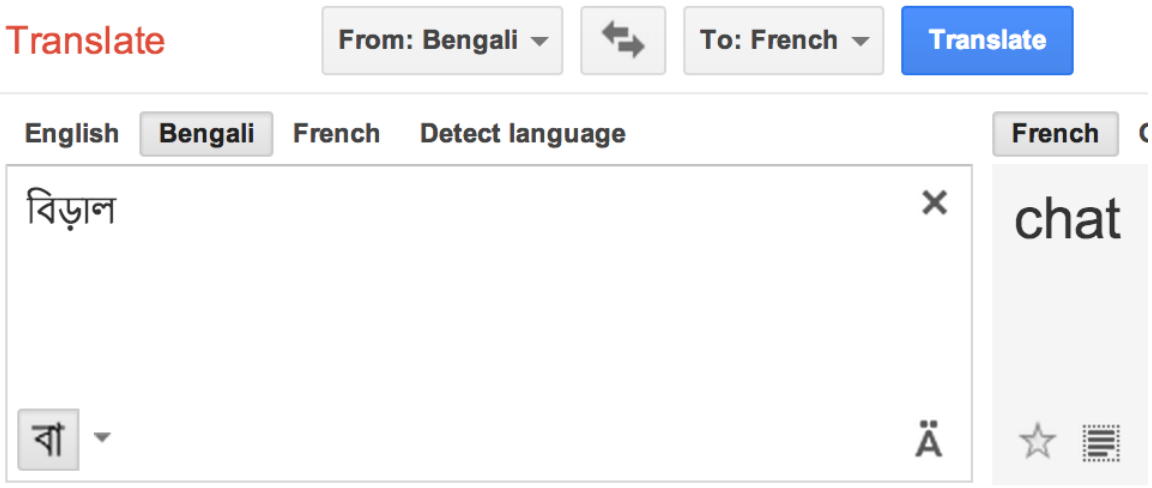
Now supports 42 languages, including accents/dialects in 46 countries



Input Tools

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





Google

Font Development at Google

We want to remove tofu on the Web!

Pan-Unicode Font (Noto = No Tofu)

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

Sfntly Font Library

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

Supports

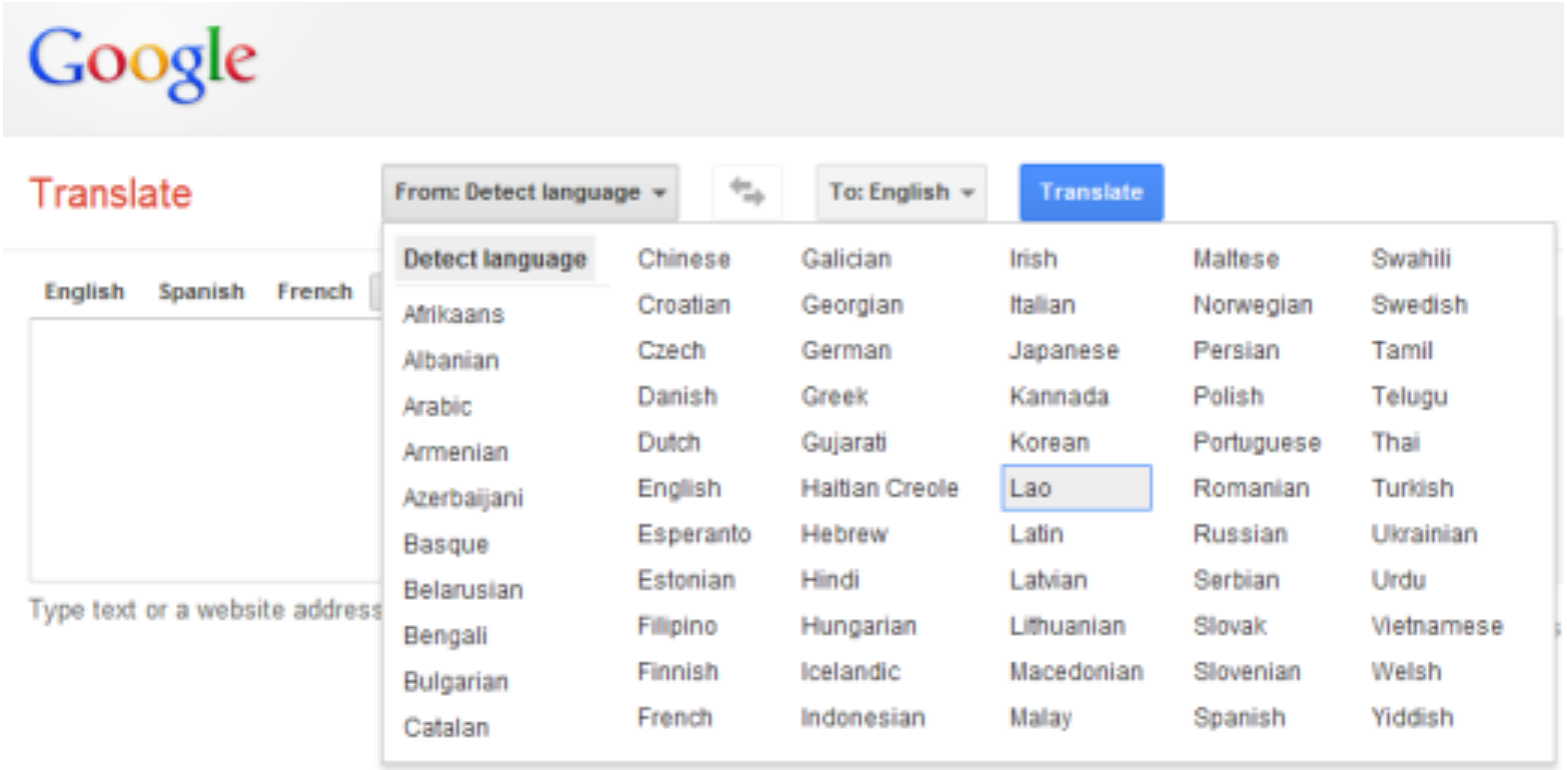
- reading and writing basic font tables
- support for bitmap glyphs
- e.g. can use to subset fonts, find problems, extract information

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

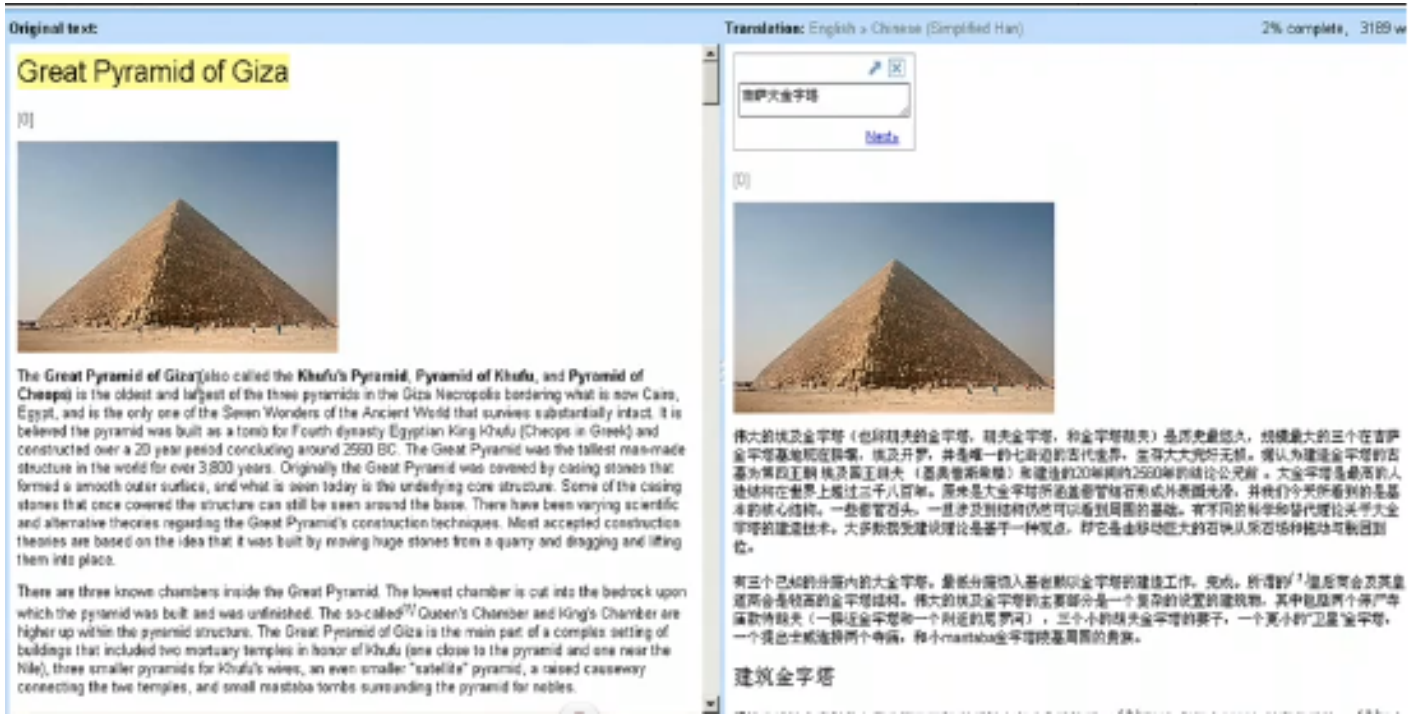


Google

Google Localization Infrastructure

http://translate.google.com/toolkit

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



Google

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.",…
```

YouTube Caption Translation

And even more...

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

Questions?