Bridging languages, cultures, and technology

26, October 2010 Peter Constable, Sr. Program Manager Jan Anders Nelson, Sr. Program Manager

Microsoft Corporation



The MultiLingualWeb

Where We Are

"Axemotivage about the partoispace axes of ae thrate the government of the second of the s Reality of the state of the sta Languages – Benjamin B. Sargent 20 February 2008,

Language Death, Watchtower

The MultiLingualWeb

Where We Are

- Microsoft Research
 - Microsoft Translator APIs & Collaborative
 Translations Framework
 - Statistical Machine Translation based cloud service bridging language gaps for millions of users
 - Currently supporting 35 languages and powering translations in many scenarios via Office, Internet Explorer and Bing
 - Translation, Language Detection, Speech and Collaborative editing public APIs
 - http://www.microsofttranslator.com

The MultiLingualWeb

Where We Are

- Microsoft Research
 - WikiBhasha Beta
 - Browser-based multilingual content creator for Wikipedia
 - Open source, integrated into Mediawiki
 - Support for set of languages powered by Microsoft Translator service (currently 35)
 - Built on Microsoft's <u>Machine Translation</u> system and <u>Collaborative Translation Framework</u>
 - http://www.wikibhasha.org/

The MultiLingualWeb Where We Are

- Microsoft Local Language Program
 - Global initiative that fosters the development and proliferation of regional language groups, enabling them to preserve and promote their language and culture while benefiting from continuing IT advancements
 - Support for 95 languages reaching 1 billion people
 - http://www.microsoft.com/LLP

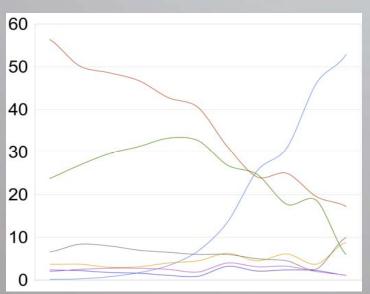
The MultiLingualWeb Where We Are

- Number of Internet users to surpass 2 billion by end of year
 - The number of Internet users worldwide has doubled in the past five years and will surpass the 2 billion mark in 2010, with the majority of the 226 million new users this year coming from developing countries
 - UN News Centre, 19 October 2010

- IE localized in 95 languages
- But... browser localization isn't what enables multilingual Web content / apps



- Browser / plug-ins / host client OS functionality
 - Text display, input, formatting...
- Global character sets
 - Unicode!
 - UTF-8 > 50%





- Global-ready HTML / CSS
- Content / app localization
- Client-server interaction
 - Charset
 - User context:
 - Language, formatting preferences
 - Current vs. home location



- Scripting
 - Interaction with host browser/OS
 - User language preferences
 - User formatting, collation preferences
 - Location (home, current)
 - Current input language
 - Linguistic services: spelling, hyphenation, ...



Current state of the art: very mixed



demos

- Script support
- International text in Canvas, SVG
- CSS: control of fonts for international text

Key challenges

- Content / web app localization
 - i18n 101: separation of code & resources
- Client-server interaction
 - Charsets still a lingering nuisance
 - Migrate to UTF-8
 - accept-language, IP heuristics: not enough!
 - Unicode locale tags (BCP 47 with -u extension)
- Scripting: limited g11n support
 - Work starting to create an ECMA standard g11n library for Javascript

questions?



Microsoft®

Your potential. Our passion.™