Building Partnerships with Language Communities: The Importance of Shared Technology and Shared Data

Bill Dolan

Microsoft Research

November 18, 2010

Outline

- Introduction
- Why partner?
- Data Scarcity
- An Experiment in Latvia
- Data Crowdsourcing
    - Community Translation Foundation
    - WikiBasha

Microsoft Translator Translation service

- State of the art Statistical Machine Translation system available as a cloud service
  - Powers millions of translations every day – in Office, Internet Explorer, Bing…
  - 35 languages and counting…
  - Constant improvements in languages and quality
  - Available to end users at microsofttranslator.com
  - Broad set of APIs and user controls for easy integration into any scenario – web, desktop or mobile
- Team sits within MSR: success is measured by academic/community impact, not just business impact

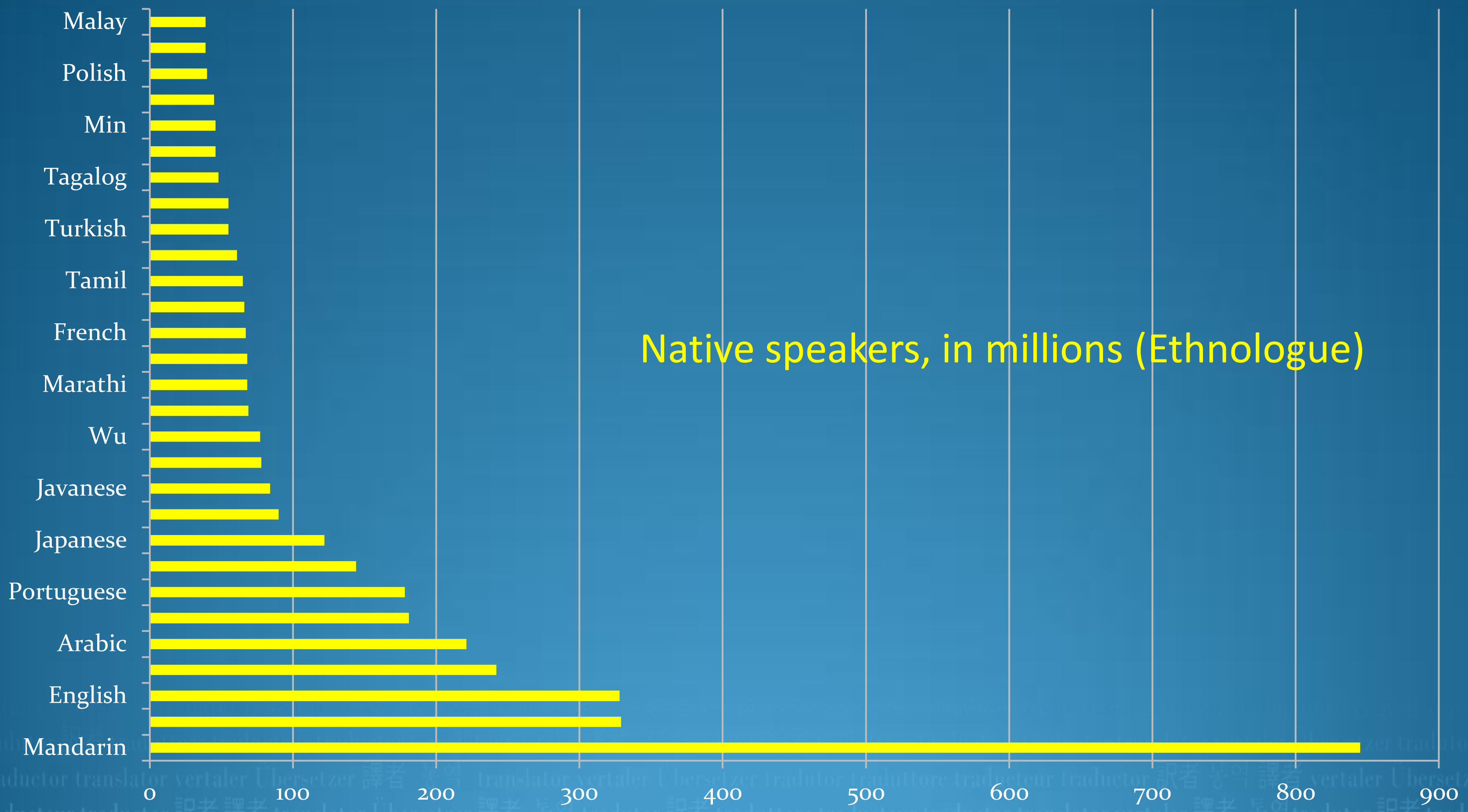
Outline

- Introduction
- Why partner?
- Data Scarcity
- An Experiment in Latvia
- Data Crowdsourcing
    - Community Translation Foundation
    - WikiBasha

How many pairs can reach “high-quality”?

- The goal is metaphorically grand:
  - “Eliminating Language Barriers”
  - “Leveling the Global Playing Field”
  - “Flattening the world”
- But how much topographical remodeling can we really do?
  - In practical terms, the scale of the problem is enormous
  - Too many languages, too many pairs, too little data
  - No matter how big your group, it’s not big enough
- The monolithic development model breaks down fast
  - Distributed development is the only model that makes sense
  - Broad-scale international collaboration is needed: corporate, academic, government, and language communities

Most of the world is going to be left out



Native speakers, in millions (Ethnologue)

- Not much data/research for e.g. English-Estonian, English-Tamil, English-Polish
  - And none for e.g. Estonian-Mandarin, Spanish-Polish, Vietnamese-Bengali

A World without Language Barriers

- No language has supremacy over others
  - Everyone speaks and writes in their native language, translation occurs seamlessly
- A Language-Neutral Natural User Interface
  - Search and browse the web without caring about the content's language origin
  - Control your car, cell phone, games, television, house, etc. using your native tongue

Microsoft® Translator

A great vision!

- But only if you speak a G20 language
- And it had better be a dominant one in your region

MT is a transformative Technology

- But its benefits are not uniformly accessible
  - As quality/usage grow, it could actually reinforce language barriers
- New economic opportunities if you speak German or French
  - No need to be bilingual
- But that’s not true if you’re a monolingual Hungarian speaker

Are we helping create a linguistically disenfranchised underclass?

So who’s to blame? Who can we sue?

- No one
  - There really isn’t a bad guy in this
- Hard for companies to justify investment in smaller markets
  - Localizing language technologies can be hugely expensive
  - If incremental costs are low, maybe “check-box” quality
- Academics have essentially the same problems
  - No resources, no time, not enough bodies, not enough data
- We all believe that NL technology is a positive force
  - But we can’t forget about low-resource languages
  - We don’t want to end up creating the very barrier we’re trying to knock down

Beyond Translation

- Investment in MT has important spillover effects on other tools and capabilities
  - LM techniques, parsers, morphological analyzers, etc.
  - Training/test corpora for spellers, input method editors, speech recognition, text-to-speech, etc.
- NUI, and speech-driven interfaces are coming fast
  - Mobile, interactive voice response systems, Kinect, Siri
  - Burnistoun video

What can we do to ensure smaller languages aren't excluded from this future?

Haitian Creole: a collaborative story

(or How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes)

- Haitian is an extremely resource-poor language
  - No corpora, no significant Web presence, idiosyncratic formats for what did exist, not a lot of easily discoverable data
- Much of the data had to be discovered manually
  - Lots of volunteer help!
- NLP community started sharing data
  - Carnegie Mellon University, CrisisCommons, Mission 4636, Ushahidi
- Companies volunteered to manually translate more
  - Butler Hill Group, WeLocalize, Moravia Worldwide
- Targeted content relevant to relief effort
- Giving back to the community through data donations
  - Data with clear license -> TAUS Data Association

But in the general case: Sharing

- Interface Standards: how does an app communicate with an MT service?
  - Dictionaries
  - Custom training data
  - Domain taxonomy
  - Security settings
  - TM upload/download
  - Any metadata returned from the service to the application
- Tools
- Data

Outline

- Introduction
- Why partner?
- Data Scarcity
- An Experiment in Latvia
- Data Crowdsourcing
    - Community Translation Foundation
    - WikiBasha

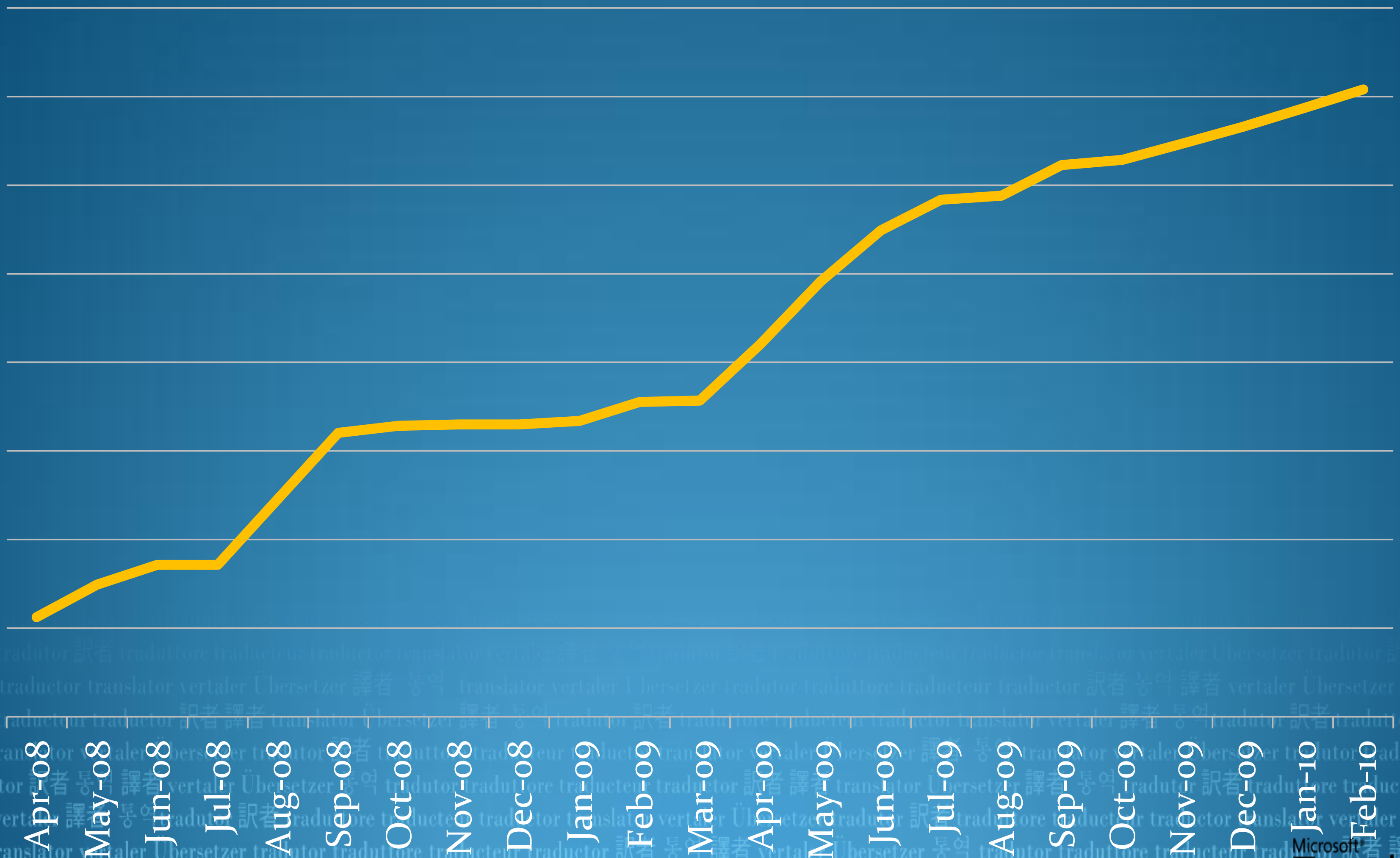
Standard Procedure Data Gathering

- Web data gathering
  - Web-scale algorithms to find parallel pages
  - Page and sentence alignment
- Existing (mostly) parallel data
  - Microsoft manuals and software
  - Dictionaries, phrasebooks
  - Government Data
  - Data sharing associations
    - Linguistic Data Consortium, Taus Data Association, ELRA, …
  - Licensed data
    - Microsoft Press, …
- Comparable (non-parallel) data
  - Wikipedia
  - News articles

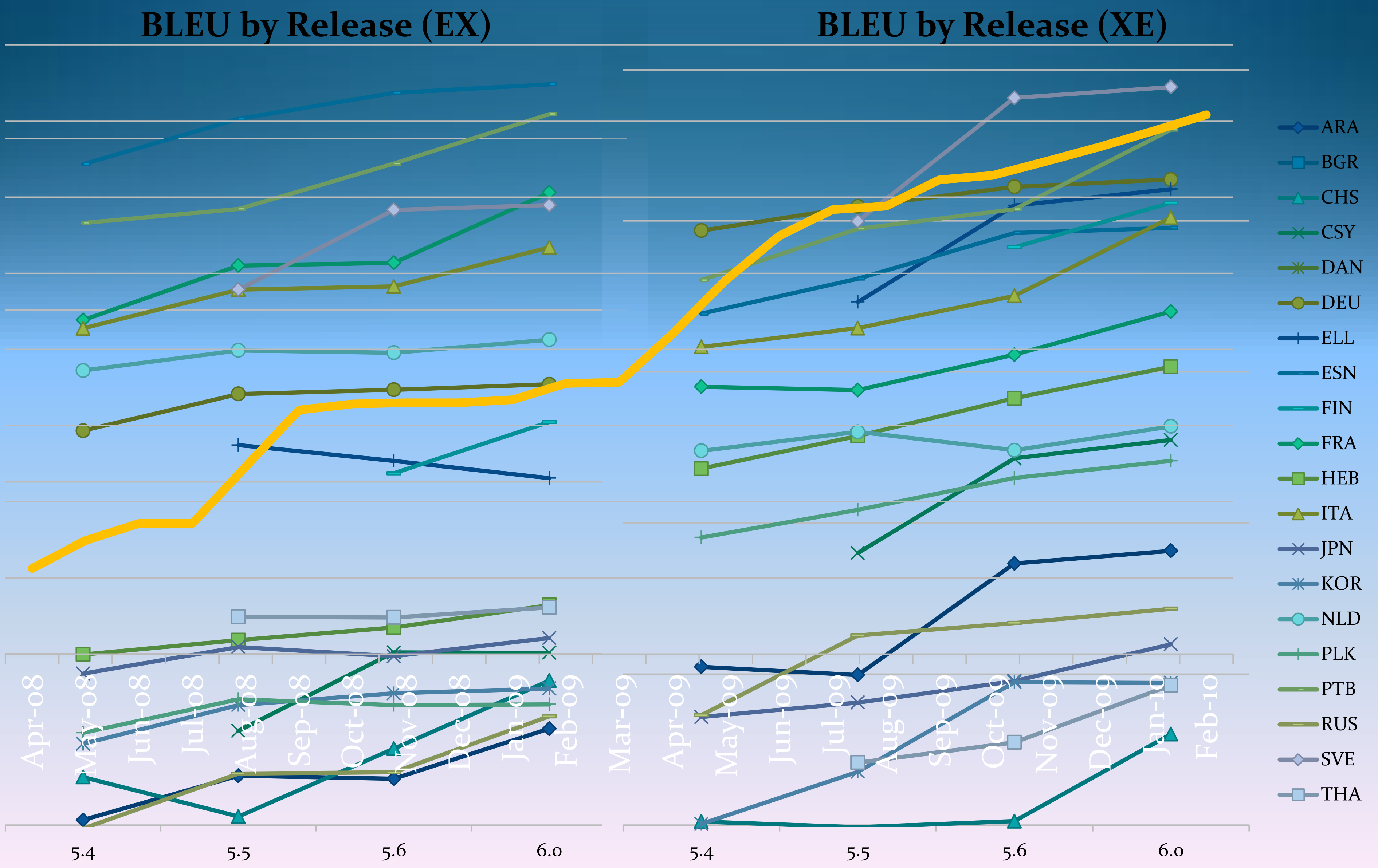
Internal Use:
Customized using mostly Microsoft and TAUS data, optimized for Microsoft content

Data volume directly affects MT quality!

Parallel Sentences



Quality improvements in 2009



BLEU by Release (EX)

BLEU by Release (XE)

Data Sources

- Web data gathering
  - Web-scale algorithms to find parallel pages
  - Page and sentence alignment
- Existing (mostly) parallel data
  - Microsoft manuals
  - ion, ELRA, …
    - Microsoft Press
- Comparable (no
  - Wikipedia
  - News articles

This is not enough!

We need more data!

And for low-resource languages we need even more!

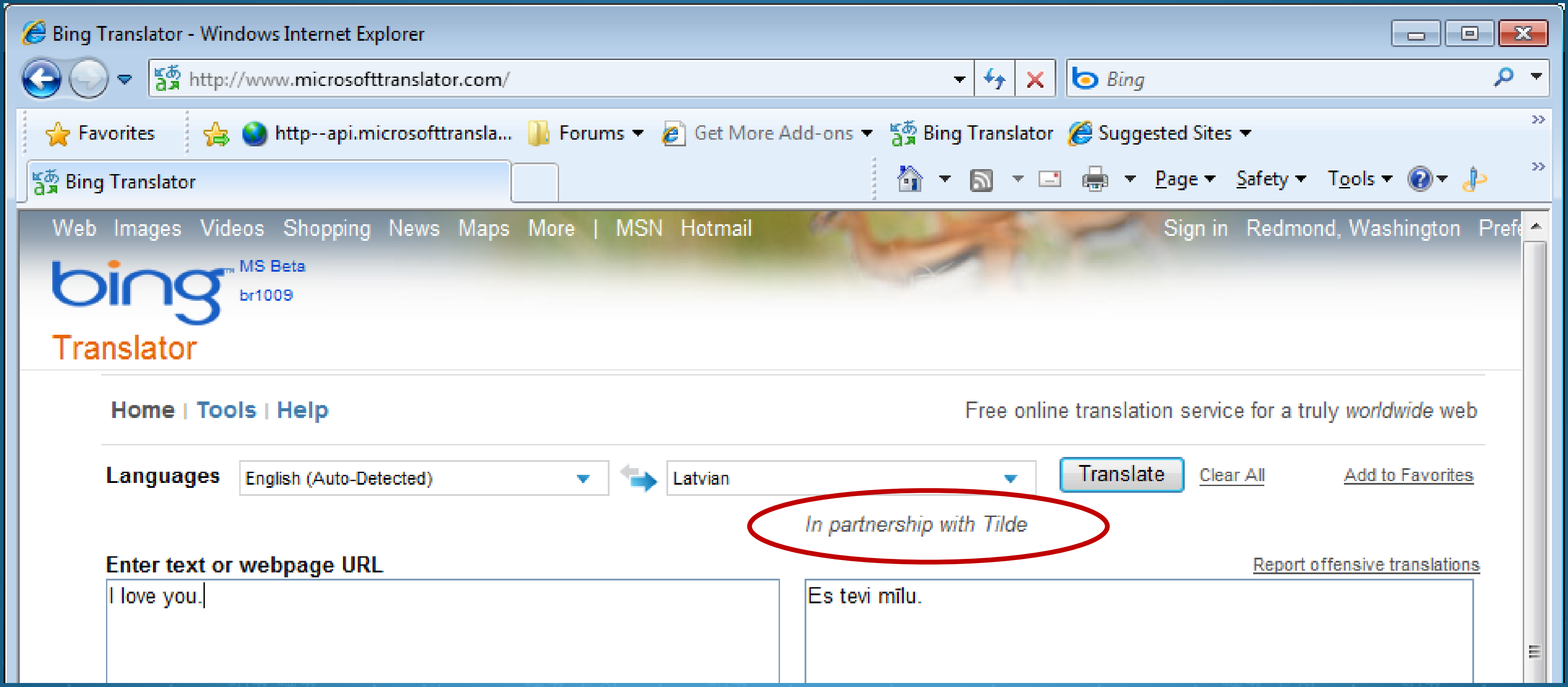
Building MT for “G21+” Languages

- Local communities must be enlisted to help
  - Both on the technical and data collection fronts
  - Who cares most about a language? Who speaks it?!
- Data is the key
  - Without it, local R&D can’t even begin
  - Publishing opportunities, progress depend on large common datasets
- We must work together—and with local communities--to build large, shared parallel datasets
  - Free of licensing issues
  - Shared through e.g. TDA or ELRA
  - Ideally, domain-classified

Outline

- Introduction
- Why partner?
- Data Scarcity
- An Experiment in Latvia
- Data Crowdsourcing
  - Community Translation Foundation
  - WikiBasha

Latvian: Collaborating with Tilde



Free online translation service for a truly *worldwide* web

Languages: English (Auto-Detected) → Latvian

In partnership with Tilde

Enter text or webpage URL

I love you.

Es tevi mīlu.

Tilde's work is directly used by Latvian users of Office, Internet Explorer, etc.

Microsoft Translator

Direct Collaboration Model

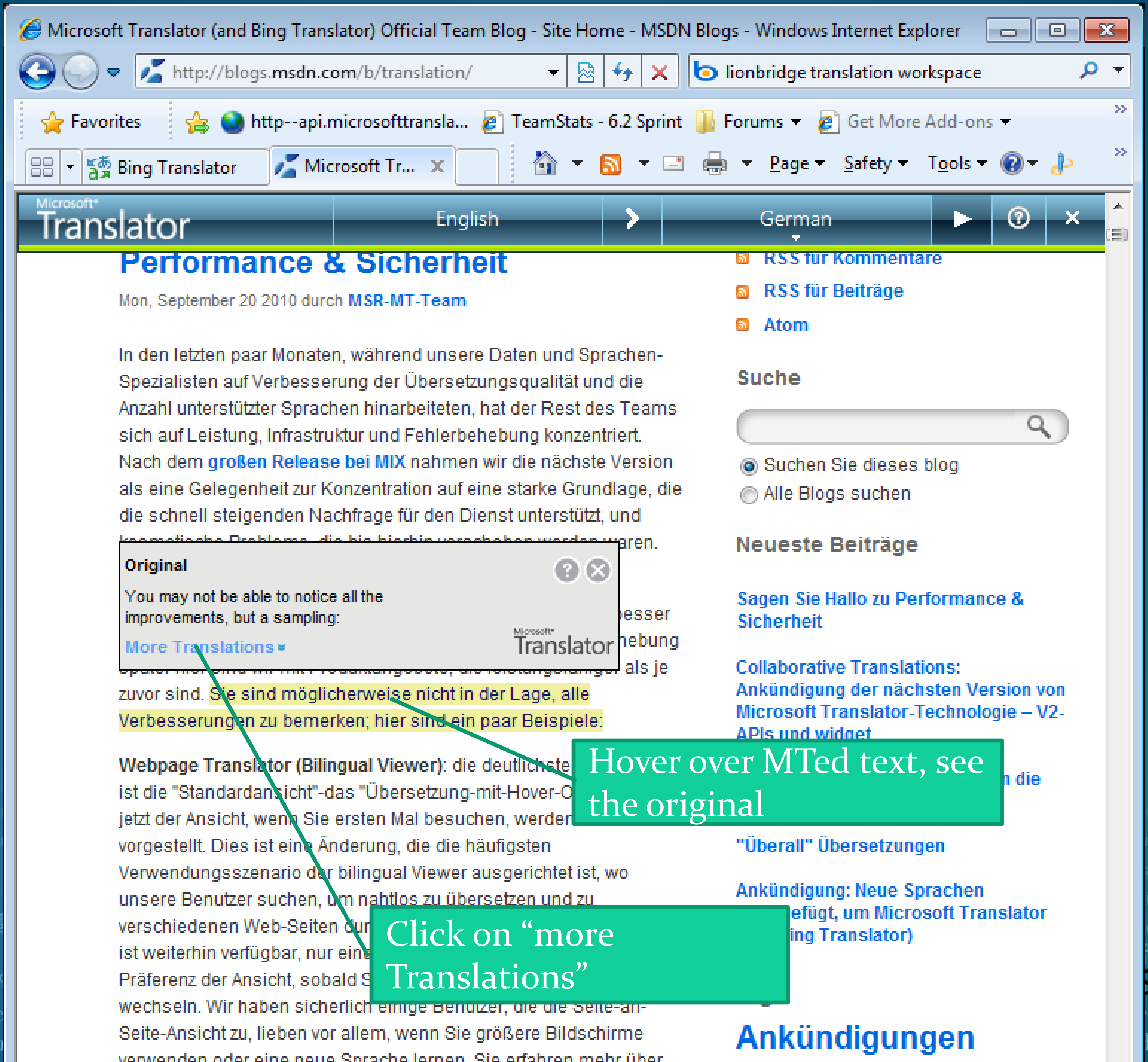
- Tilde's skilled developers worked directly with MSR team to:
  - Incorporate Latvian morphological processing
  - Build, test, and deploy models on http://microsofttranslator.com
- Data Sharing
  - Tilde's connections allowed it to identify significant amounts of parallel data that wasn't on the web
  - MSR and Tilde shared data when legally possible
- A win-win-win-win-win: public/private partnership
  - Mindshare for Tilde via exposure in MS Office, better Latvian-English MT for MS
  - The Latvian government is happy
  - The Latvian language and NL research communities have a growing public data resource, new awareness of NL technology's importance

Outline

- Introduction
- Why partner?
- Data Scarcity
- An Experiment in Latvia
- Data Crowdsourcing
    - Community Translation Foundation
    - WikiBasha

Crowdsourcing in Latvia

- Tilde coordinated a local crowd-sourced data collection effort
- Collaborative Translation Framework (CTF)
    - MT post-editing scenario, in-place on your web site
    - Collects votes, feedback and corrections from users of deployed machine translation
    - Enables the content owner to approve the corrections, or delegate the approval authority to others.


Microsoft Translator (and Bing Translator) Official Team Blog - Site Home - MSDN Blogs - Windows Internet Explorer
http://blogs.msdn.com/b/translation/
lionbridge translation workspace
Favorites
http--api.microsofttransla...
TeamStats - 6.2 Sprint
Forums
Get More Add-ons
Bing Translator
Microsoft Tr...
Page
Safety
Tools
Microsoft Translator
English
German

Performance & Sicherheit

Mon, September 20 2010 durch MSR-MT-Team

In den letzten paar Monaten, während unsere Daten und Sprachen-Spezialisten auf Verbesserung der Übersetzungsqualität und die Anzahl unterstützter Sprachen hinarbeiteten, hat der Rest des Teams sich auf Leistung, Infrastruktur und Fehlerbehebung konzentriert. Nach dem großen Release bei MIX nahmen wir die nächste Version als eine Gelegenheit zur Konzentration auf eine starke Grundlage, die die schnell steigenden Nachfrage für den Dienst unterstützt, und kosmetische Probleme, die bis hierhin verschoben worden waren.

Original

You may not be able to notice all the improvements, but a sampling:

More Translations

Microsoft Translator

...besser ...hebung ... als je zuvor sind. Sie sind möglicherweise nicht in der Lage, alle Verbesserungen zu bemerken; hier sind ein paar Beispiele:

Webpage Translator (Bilingual Viewer): die deutlichste ... ist die "Standardansicht"-das "Übersetzung-mit-Hover-O... jetzt der Ansicht, wenn Sie ersten Mal besuchen, werden ... vorgestellt. Dies ist eine Änderung, die die häufigsten Verwendungsszenario der bilingual Viewer ausgerichtet ist, wo unsere Benutzer suchen, um nahtlos zu übersetzen und zu verschiedenen Web-Seiten ... ist weiterhin verfügbar, nur eine ... Präferenz der Ansicht, sobald S... wechseln. Wir haben sicherlich einige Benutzer, die die Seite-an-Seite-Ansicht zu, lieben vor allem, wenn Sie größere Bildschirme verwenden oder eine neue Sprache lernen. Sie erfahren mehr über

RSS für Kommentare

RSS für Beiträge

Atom

Suche

● Suchen Sie dieses blog

○ Alle Blogs suchen

Neueste Beiträge

Sagen Sie Hallo zu Performance & Sicherheit

Collaborative Translations: Ankündigung der nächsten Version von Microsoft Translator-Technologie – V2-APIs und widget

... die

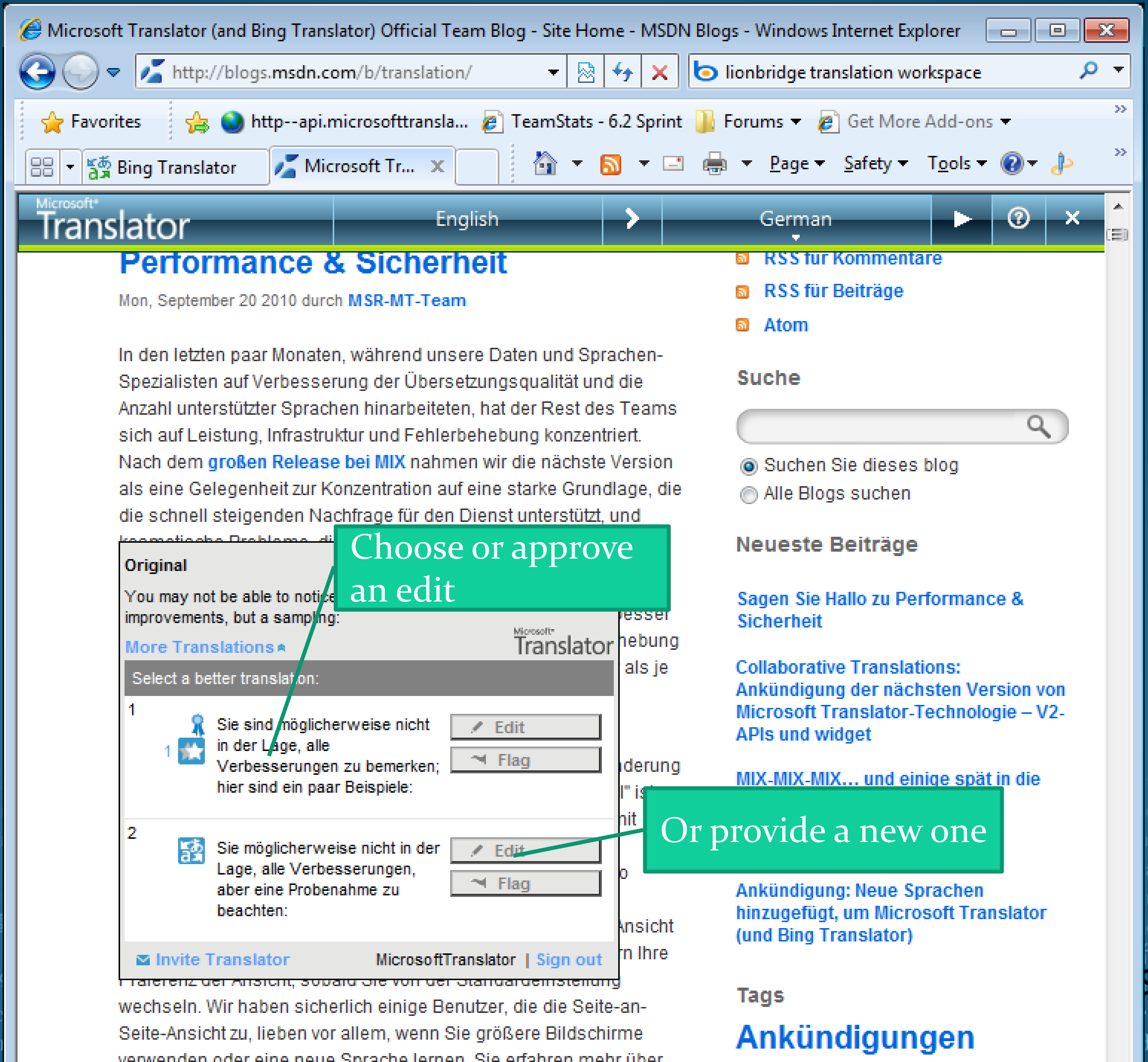
"Überall" Übersetzungen

Ankündigung: Neue Sprachen ...efügt, um Microsoft Translator ...ing Translator)

Ankündigungen

Hover over MTed text, see the original

Click on "more Translations"


Microsoft Translator (and Bing Translator) Official Team Blog - Site Home - MSDN Blogs - Windows Internet Explorer
http://blogs.msdn.com/b/translation/
lionbridge translation workspace
Favorites
http--api.microsofttransla...
TeamStats - 6.2 Sprint
Forums
Get More Add-ons
Bing Translator
Microsoft Tr...
Page
Safety
Tools
Microsoft Translator
English
German

Performance & Sicherheit

Mon, September 20 2010 durch **MSR-MT-Team**

In den letzten paar Monaten, während unsere Daten und Sprachen-Spezialisten auf Verbesserung der Übersetzungsqualität und die Anzahl unterstützter Sprachen hinarbeiteten, hat der Rest des Teams sich auf Leistung, Infrastruktur und Fehlerbehebung konzentriert. Nach dem **großen Release bei MIX** nahmen wir die nächste Version als eine Gelegenheit zur Konzentration auf eine starke Grundlage, die die schnell steigenden Nachfrage für den Dienst unterstützt, und kosmetische Probleme, di

besser
hebung
als je

derung
l" is
nit
o

Ansicht
n Ihre
Präferenz der Ansicht, sobald Sie von der Standardeinstellung wechseln. Wir haben sicherlich einige Benutzer, die die Seite-an-Seite-Ansicht zu, lieben vor allem, wenn Sie größere Bildschirme verwenden oder eine neue Sprache lernen. Sie erfahren mehr über

Original

You may not be able to notice improvements, but a sampling:

More Translations

Microsoft Translator

Select a better translation:

1 Sie sind möglicherweise nicht in der Lage, alle Verbesserungen zu bemerken; hier sind ein paar Beispiele:

Edit
Flag

2 Sie möglicherweise nicht in der Lage, alle Verbesserungen, aber eine Probenahme zu beachten:

Edit
Flag

Invite Translator | MicrosoftTranslator | Sign out

RSS für Kommentare
RSS für Beiträge
Atom

Suche

● Suchen Sie dieses blog
○ Alle Blogs suchen

Neueste Beiträge

Sagen Sie Hallo zu Performance & Sicherheit

Collaborative Translations: Ankündigung der nächsten Version von Microsoft Translator-Technologie – V2-APIs und widget

MIX-MIX-MIX... und einige spät in die

Ankündigung: Neue Sprachen hinzugefügt, um Microsoft Translator (und Bing Translator)

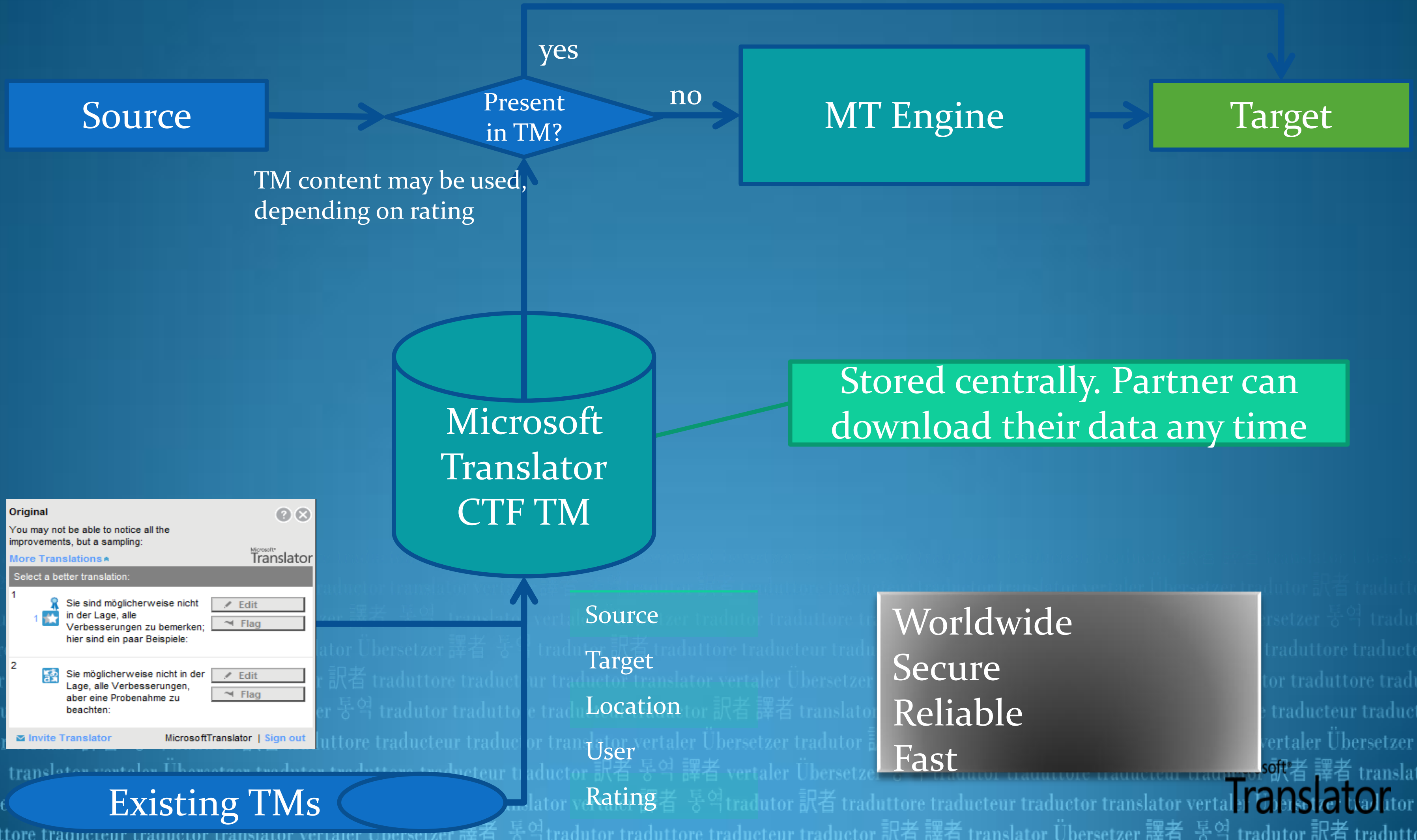
Tags

Ankündigungen

Choose or approve an edit

Or provide a new one

Collaborative Translations Framework (CTF)



Source

Present in TM?

yes

no

MT Engine

Target

TM content may be used, depending on rating

Microsoft Translator CTF TM

Stored centrally. Partner can download their data any time

Original

You may not be able to notice all the improvements, but a sampling:

More Translations

Select a better translation:

1 Sie sind möglicherweise nicht in der Lage, alle Verbesserungen zu bemerken; hier sind ein paar Beispiele:

2 Sie möglicherweise nicht in der Lage, alle Verbesserungen, aber eine Probenahme zu beachten:

Existing TMs

Source
Target
Location
User
Rating
…

Worldwide
Secure
Reliable
Fast

CTF available through a set of APIs

http://sdk.microsofttranslator.com/

- Fully integrated into the Microsoft Translator API set
- Available in AJAX, SOAP and REST flavors.
- Anything submitted within your site is yours
  - Download freely

Motivating the Latvian Crowd

- Many organized activities
  - 700+ people heard the message, >6K participants in 2 months
  - Public discussion organised in co-operation with the National Library of Latvia, live broadcast on internet
  - Presentation to the representatives of regional libraries
  - *E-seminar* presentation
  - Presentation at *BarCamp* 2010 ‘unconference’ / ‘mashup’
- Tilde presented the effort to the Latvian public as:
  - For or the common good: developing technological support for the Latvian language
  - A scientific, rather than commercial, effort
  - Emphasized that data would be shared back to community
- The president of Latvia publicly supported the effort

Outline

- Introduction
- Why partner?
- Data Scarcity
- An Experiment in Latvia
- Data Crowdsourcing
  - Community Translation Foundation
  - WikiBasha

WikiBhasha

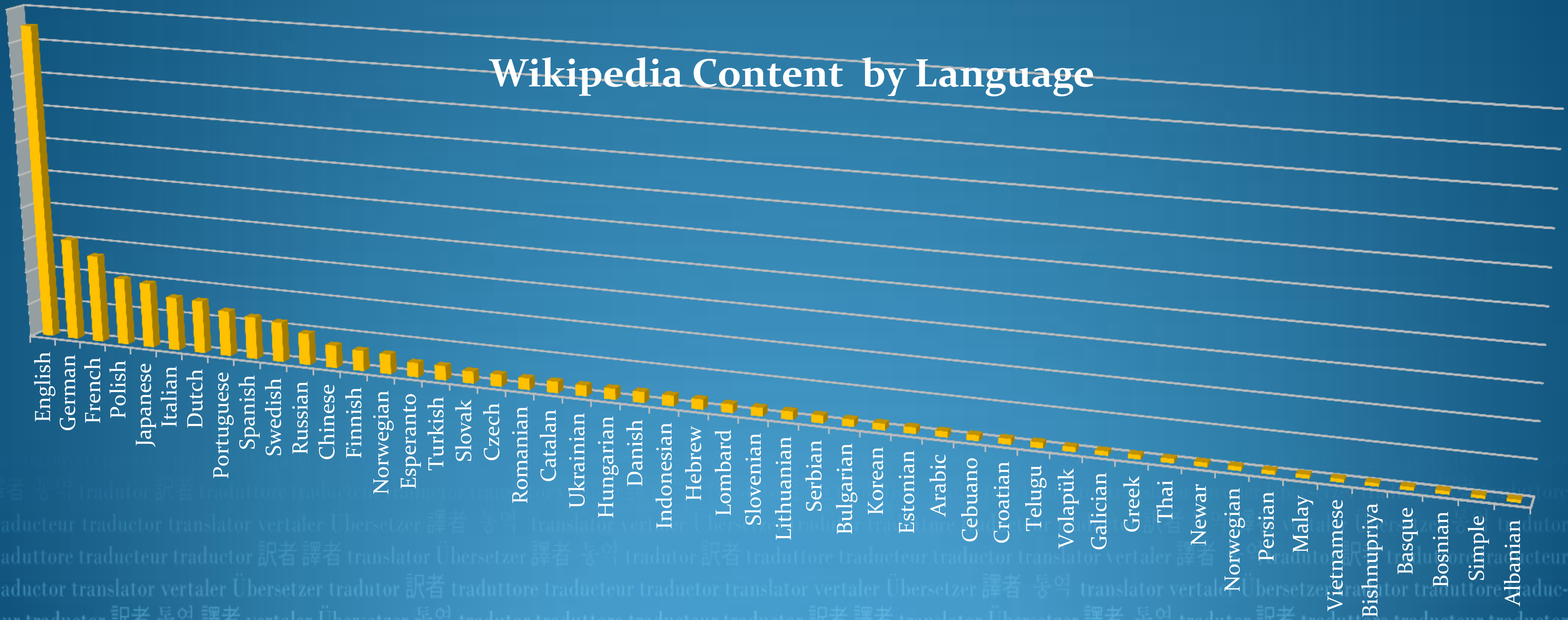
“Wiki” + “Bhasha” (“*language*” in Hindi & Sanskrit)

- An application that exploits the CTF API
- Code released as an open-source Media Wiki extension
- A browser-based application on Wikipedia
  - Helps users create multilingual content in non-English Wikipedias
  - Targets low-resource languages
  - Simultaneously creates useful local content + bilingual data

Demo

WikiBhasha: Why?

- Wikipedia is hugely English-biased



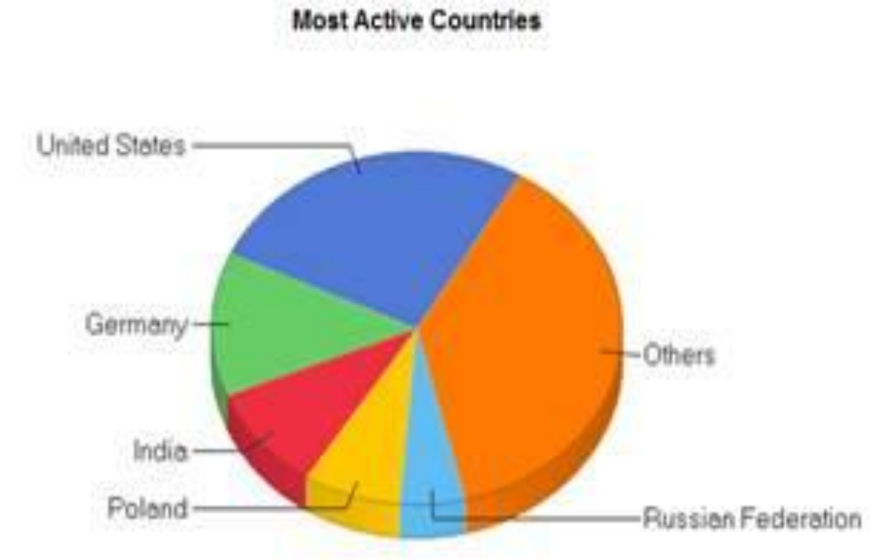
Wikipedia Content by Language

WikiBhasha now a Community Project

- Result of a formal collaboration between MSR and the WikiMedia Foundation
  - http://www.WikiBhasha.org and on Wikipedia
  - Please contribute to Wikipedia!
- WikiBhasha code
  - http://svn.wikimedia.org/viewvc/mediawiki/trunk/extensions/WikiBhasha
  - Please enhance it!

WikiBhasha: Some Statistics…

- Announced jointly by WikiMedia Foundation + MSR, October 2010
- News article covered independently in 20+ countries
- 30K Visitors, with 250K Hits in the first week
- Visitors from 50+ countries
- Hosted on Windows Azure, 99.99+ uptime



Public Website Health Status for Microsoft Research India - For public status - WikiBhasha - Windows Internet Explorer

Microsoft Research India

Current status Microsoft Research India - For public status

Current Performance and Availability Status WikiBhasha

Service is operating normally

Performance and Availability History

Oct 25	Oct 24	Oct 23	Oct 22	Oct 21	Oct 20	Oct 19
✓	✓	✓	✓	✓	✓	✓

✓ Service is operating normally ● Performance issues ● Service disruption Informational message

WikiBhasha: What next?

- Now for the hard part: motivating the crowd
  - MSR working with Wikipedia Communities around the world
- Workshops planned in several international demographics
  - India in Nov-Dec 2010
  - Egypt in Dec 2010
  - Brazil and Mexico in Jan 2011
  - Europe/Japan in 1Q 2011
- Collaborating with the Wikimedia Foundation to ensure that the data will be available as a public resource
  - Useful for MT, language modeling, etc.

Linguistic Inequalities have always been with us

- Specific languages/dialects are imbued with prestige (or not) for all kinds of historical, random reasons
  - But now we risk automating the construction of new inequalities
  - The G20 language communities get richer, the rest get poorer
- We must actively work to make sure smaller languages don't fall behind
- A monolithic approach to MT and other NL technologies will not scale. Instead,
  - Share technologies, data, agree on standards
  - Involve local governments and language communities

Thank you!

billdol@microsoft.com