LetsMT! – Online Platform for Sharing Training Data and Building User-Tailored Machine Translation

Andrejs Vasiļjevs, Tatiana Gornostay, Raivis Skadiņš

Tilde

Baltic HLT 2010, Riga, 08.10.2010

Data challenge

- **Statistical methods** provide breakthrough in cost-effective MT development
- Quality of SMT systems largely **depends on the size** of training data
- To overcome gap in SMT language and domain coverage and to improve quality much larger volume of training **data is needed**
- Parallel data accessible on the web is **just a fraction** of all translated texts. Most of them still reside in the local systems of different corporations, public and private institutions, desktops of individual users.

Let's MT!

Customization challenge

- Current mass-market and online MT systems are of **general nature** and perform poorly for domain and user specific texts.
- System adaptation is prohibitively **expensive service** not affordable to smaller companies or the majority of public institutions.
- Particulary **localization industry** is not able to fully exploit the data they have.

Let's MT!

Platform challenge

- Great open source platforms like GIZA++ and Moses make it relatively easy to build MT engine.
- Still expertise and local infrastructure is needed that is not available for majority of users.

Let's MT!

LetsMT! Vision

Let's advance MT together!

- To fully exploit the huge potential of existing open SMT technologies to create an innovative online collaborative platform for **data sharing and MT building**.
- LetsMT! is building a platform that gathers public and user-provided MT training data and generates multiple MT systems by combining and prioritizing this data.
- LetsMT! extends the use of existing state-of-the-art SMT methods that will be applied to data supplied by users to **increase quality, scope and language coverage** of machine translation.

Let's MT!

LetsMT! Vision

- Sustainable user-driven MT factory on the *cloud* providing services for user data sharing, MT generation, customization and running.

Let's MT!

LetsMT! Project ID

- Funded under: EU Information and Communication Technologies Policy Support Programme
- Area: CIP-ICT-PSP.2009.5.1 Multilingual Web: Machine translation for the multilingual web
- Project reference: 250456
- Execution: From 01/03/2010 to 31/08/2012

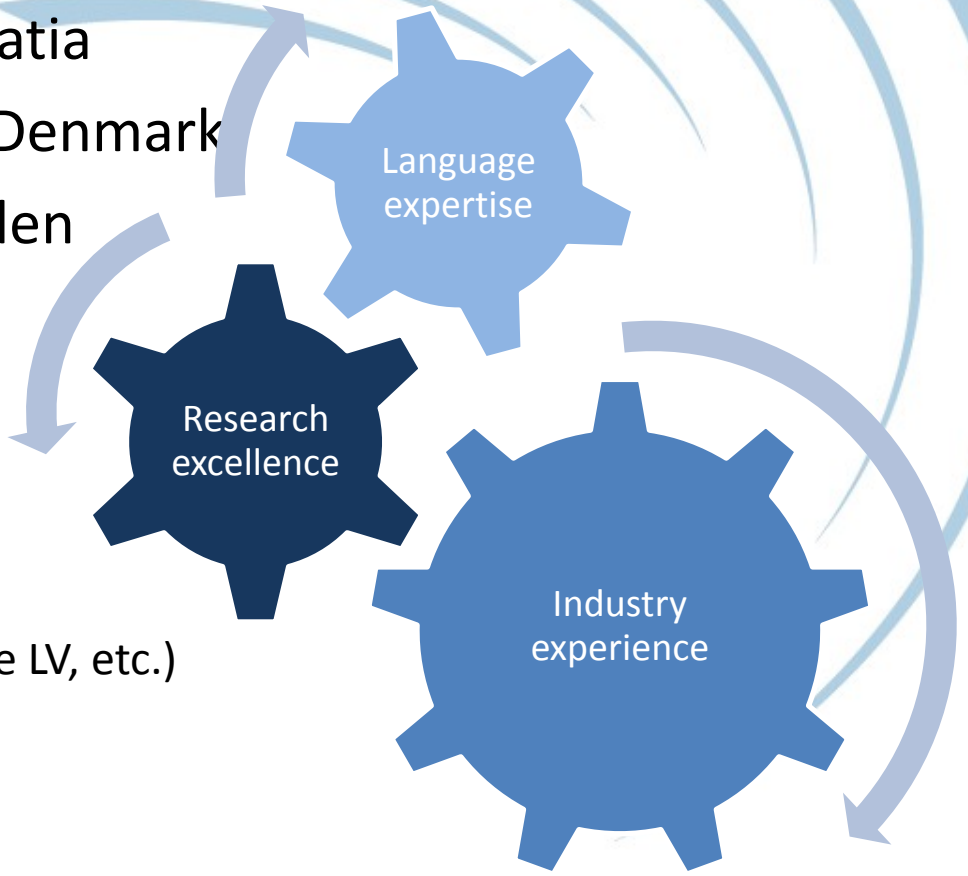
Let's MT!

ICTPSP
ICT POLICY SUPPORT PROGRAMME

Partnership with Complementing Competencies

- Tilde (Project Coordinator) - Latvia
- University of Edinburgh - UK
- University of Zagreb - Croatia
- Copenhagen University - Denmark
- Uppsala University - Sweden
- Moravia – Czech Republic
- SemLab – Netherlands

+ Support Group
(TAUS DA, SDI Media, Patent Office LV, etc.)



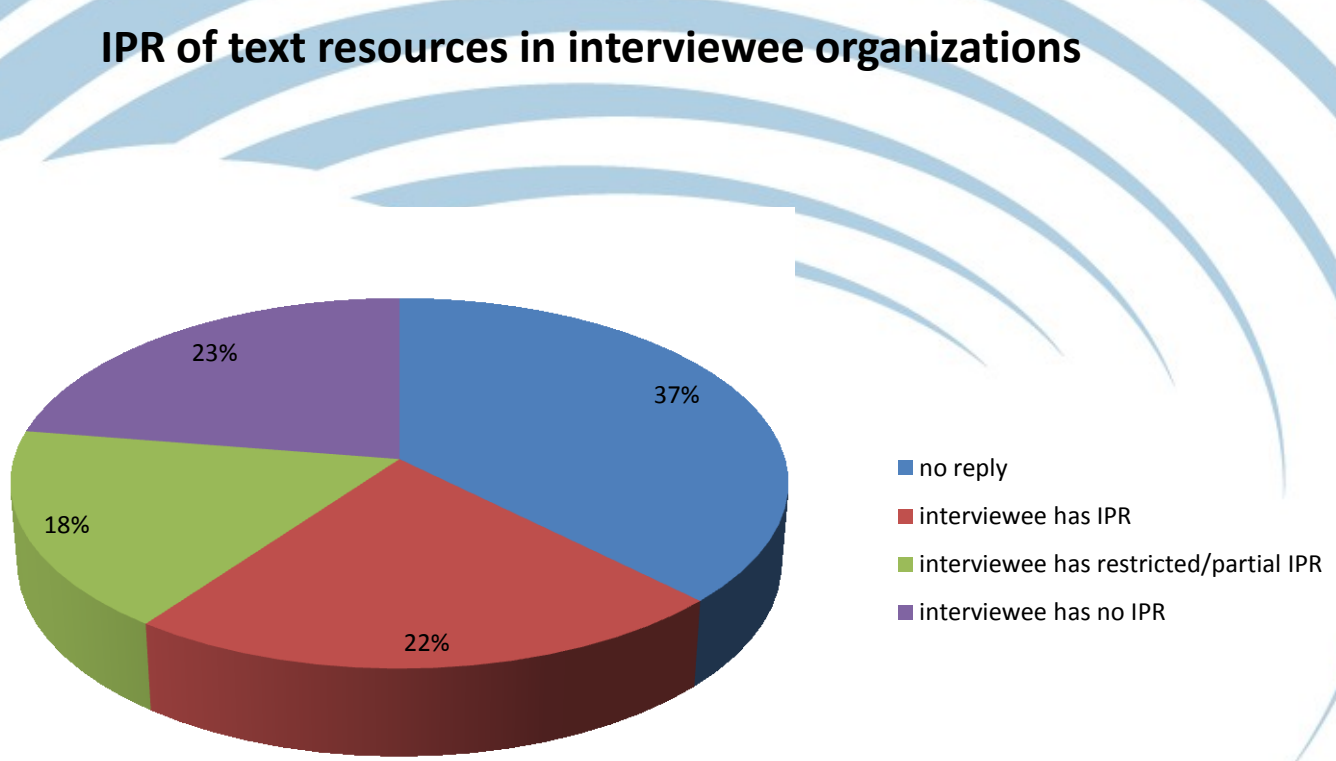
Let's MT!

LetsMT! Main Features

- Users will contribute with **user-provided content** by uploading their parallel texts
- **Directory** of web and offline resources gathered by LetsMT! as well as user provided links to other sources that are not yet included in LetsMT! repository
- **Automated training** of SMT systems from specified collections of training data
- Larger donors or customers will be able to specify particular training data collections and build **customised MT engines** from these collections
- Customers will be able to use LetsMT! platform for tailoring MT system to their needs from their **non-public data**
- Users will be involved in **MT evaluation**

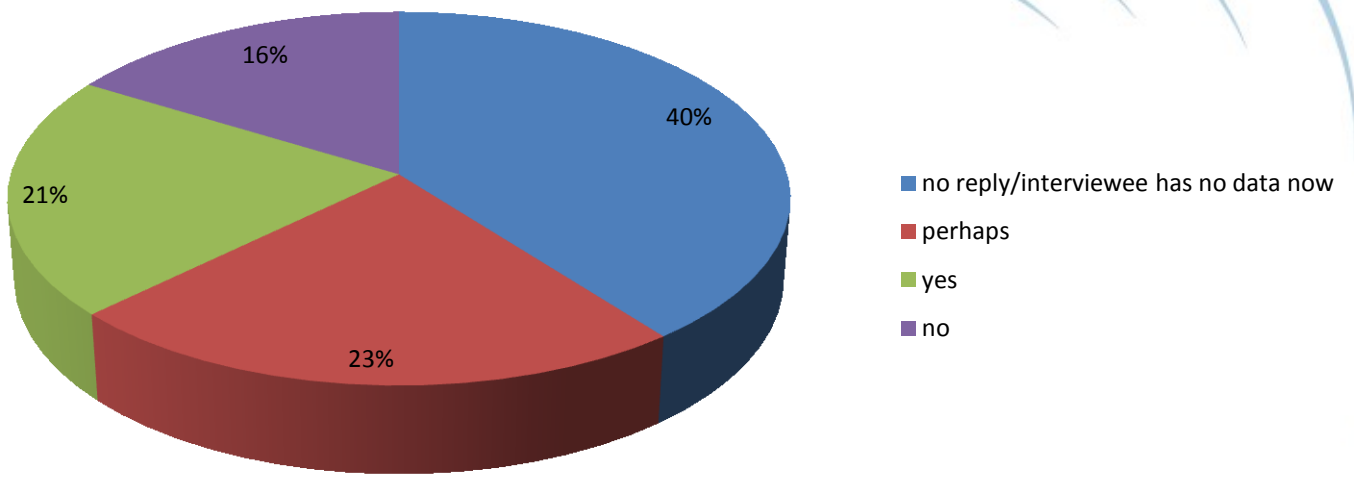
Let's MT!

User survey

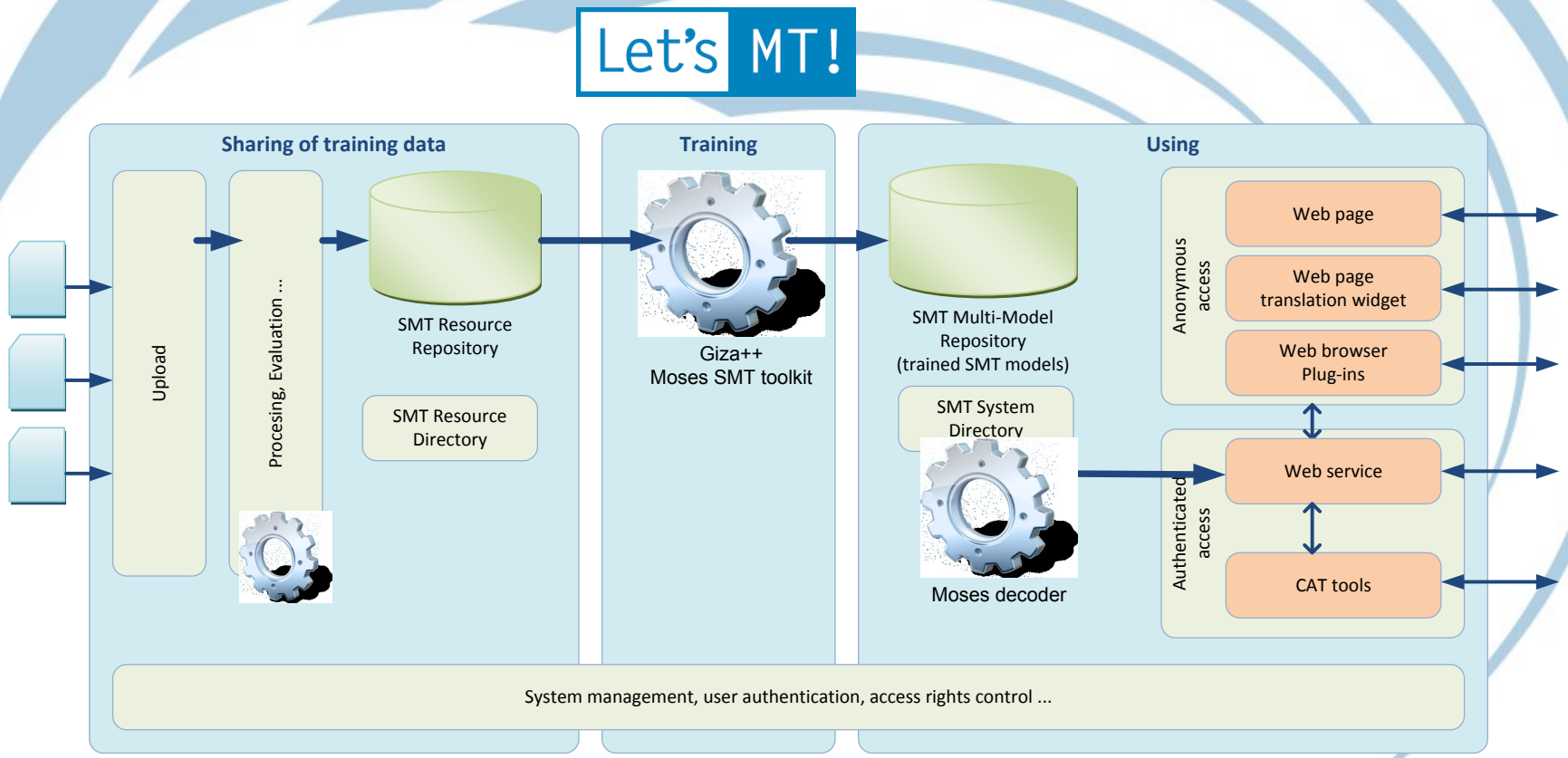


Let's MT!

User survey


Organizations' ability to share data
16%
40%
21%
23%
no reply/interviewee has no data now
perhaps
yes
no

Software Architecture



Let's MT!

Data upload formats

- Phase 1 Upload formats supported:
  - **pre-aligned** parallel corpora in standard formats - TMX, Moses format
  - **monolingual** data in popular formats DOC, PDF, plain text
- Phase 2 Upload formats supported:
  - **unaligned** parallel documents in a few popular formats DOC, PDF, plain text
  - online sentence alignment
  - possibilities to inspect intermediate results (after conversion, pre-processing, alignment)
  - possibilities to influence processing steps (adjusting parameters, selecting alternative tools that have been integrated in LetsMT!).

Let's MT!

Improvements in Moses in development by LetsMT!

- Adaptations to fit into the rapid training, updating and interactive access environment of the LetsMT! platform
- Streamlined, automatically configurable training process
- Incremental training of MT models
- Improvements in Randomized language models
- Language model and translation model servers

Let's MT!

Key Outcomes

- **website for upload** of parallel corpora and building of specific MT solutions
- **website for translation** where source text can be typed and translated
- **translation widget** provided for free inclusion into websites to translate their content
- **browser plug-ins or add-ons** that would allow the quickest access to translation
- web service for **integration in CAT tools** and other applications

Let's MT!

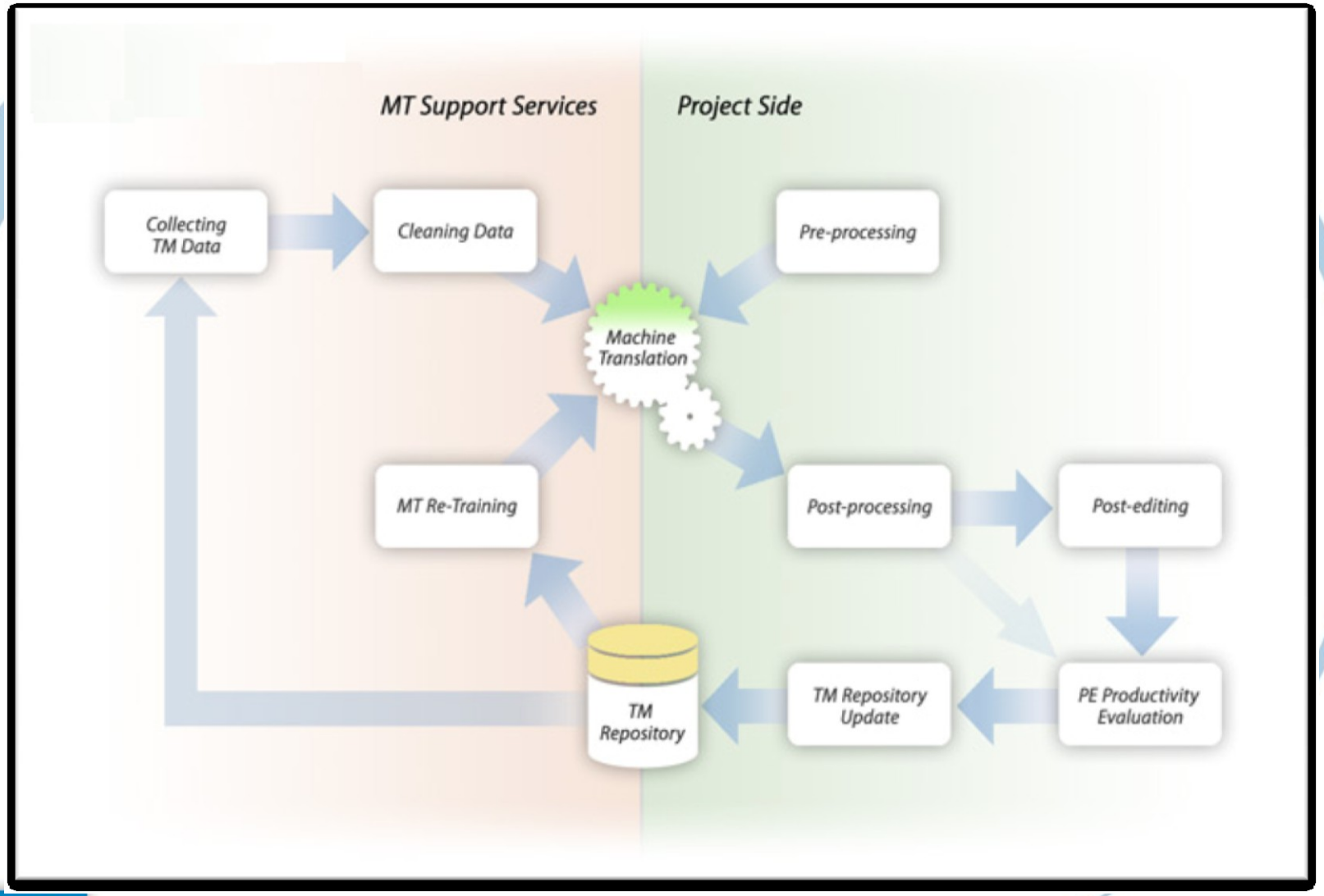
Application Scenarious

- Online MT service for the **localization and translation** industry
- Online MT service for global **business and financial news**

+ Showcase for patent translations for gisting purposes

Let's MT!

Application in localization workflow



Let's MT!

Thank you and Let’s MT!

letsmt.eu

Let’s MT!