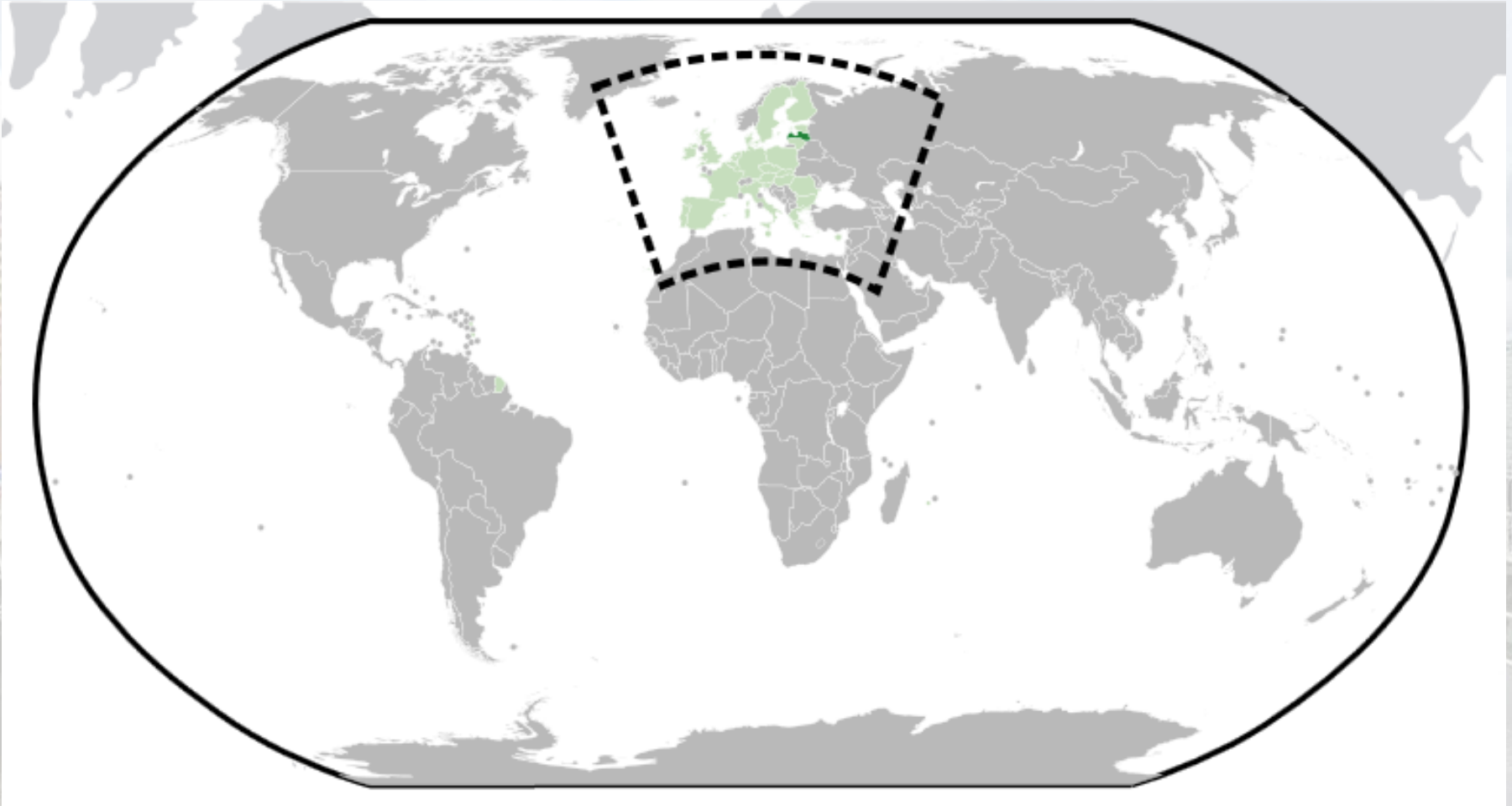


Extending the Use of Web-Based Terminology Services

Tatiana Gornostay
Tilde, Latvia

Multilingual Web Workshop, Dublin, Ireland
June 11, 2012

LATVIA



TILDE tilde.com

- Translation and Localization services
 - Latvian, Lithuanian, Estonian
- Terminology development and management
 - EuroTermBank: >2 mil terms, >25 languages
- Language Technologies and Resources
 - Small languages
- 3 offices
 - Riga (Latvia, headquarters)
 - Vilnius (Lithuania)
 - Tallinn (Estonia)
- >100 employees
 - 4 PhDs and 3 PhD candidates

European cooperation



11 June, 2012



Terminology

- Terminology is everywhere
 - visiting a doctor
 - building a house
 - buying a car
- We come across with terms every day

Terminology

- Terminology matters
 - efficient and precise communication
 - academia
 - industry
 - government

Society

Terminology

- Terminology is a language
 - Language for Specific (professional) Purposes (LSP)
 - multilingual consolidated and harmonized terminology is already being utilized as data by human users
 - language workers
 - translators, terminologists, technical writers, editors, etc.
 - now it is being developed as a web-based service for machines as users
 - systems
 - machine translation, indexing, search, annotation, etc.

Challenges

- creation, consistency, extraction
 - according to recent surveys, 84% professionals select terms from documents manually
 - acquisition
 - = term identification in a text
 - recognition
 - = term comparison with existing resources
- consolidation & harmonization
- sharing & interoperability
 - MT domain adaptation
 - concept formalization
- data annotation, indexing and search, etc.

Terminology is on the cusp between
semantic and language technologies

Terminology is bridging the three communities

Linked Open Data

Multilingual Web

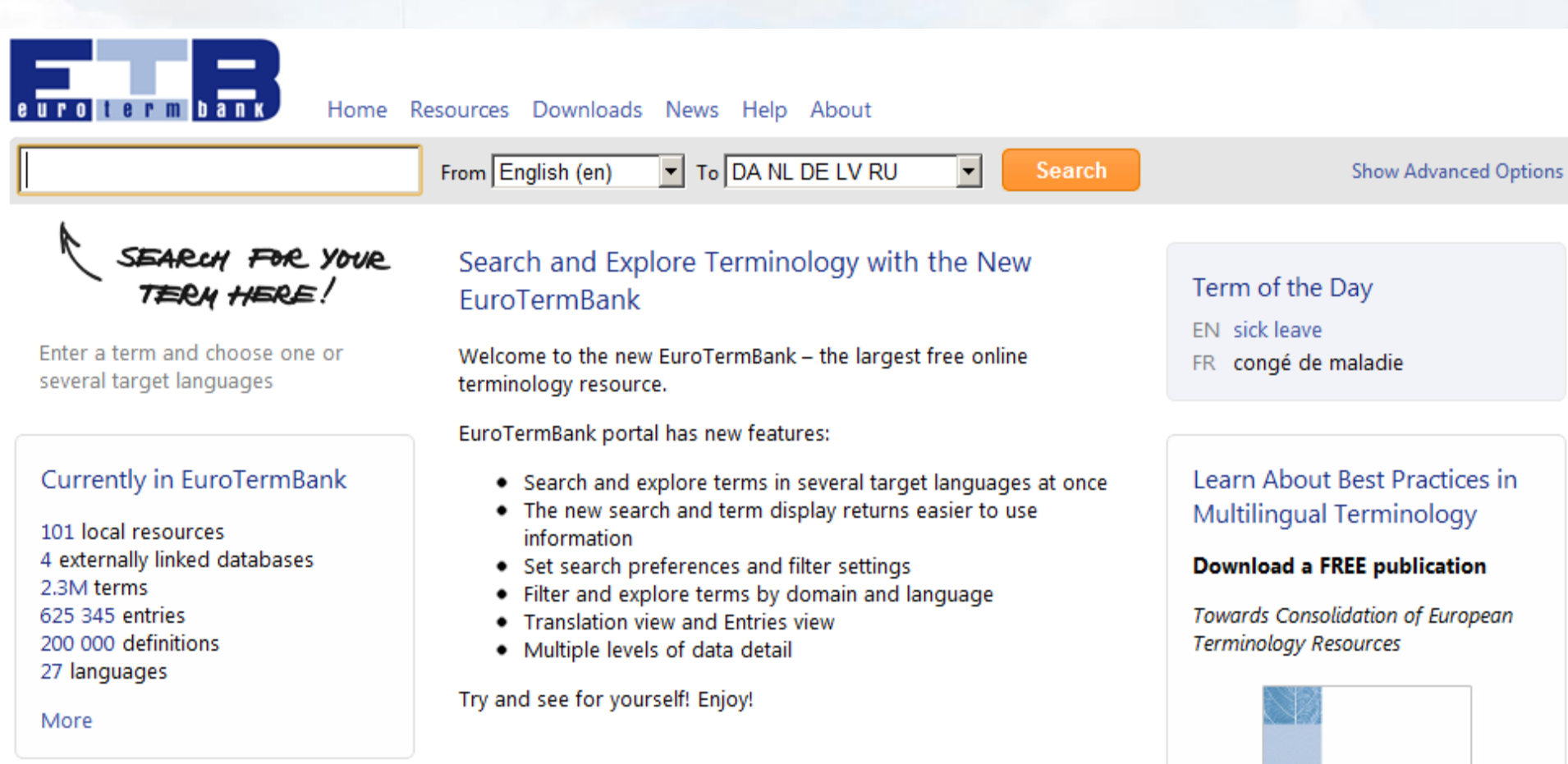
Multilingual Language Technologies, i.e. NLP

A background image showing a white, ornate bridge with a decorative railing crossing a river. In the background, there are multi-story brick buildings with many windows. The sky is blue with scattered white clouds. The entire image is slightly faded to make the text stand out.

Tilde's best practices & use cases

EuroTermBank

- www.eurotermbank.eu



The screenshot shows the EuroTermBank website. At the top left is the logo 'ETB euro term bank'. To its right is a navigation menu with links: Home, Resources, Downloads, News, Help, About. Below the navigation is a search bar with a dropdown menu set to 'English (en)' and another dropdown menu set to 'DA NL DE LV RU'. To the right of the search bar is an orange 'Search' button and a link 'Show Advanced Options'. Below the search bar, there is a handwritten note: 'SEARCH FOR YOUR TERM HERE!' with an arrow pointing to the search input field. The main content area has a heading 'Search and Explore Terminology with the New EuroTermBank' and a sub-heading 'Welcome to the new EuroTermBank – the largest free online terminology resource.' Below this is a list of features: 'EuroTermBank portal has new features:' followed by a bulleted list: 'Search and explore terms in several target languages at once', 'The new search and term display returns easier to use information', 'Set search preferences and filter settings', 'Filter and explore terms by domain and language', 'Translation view and Entries view', and 'Multiple levels of data detail'. Below the list is the text 'Try and see for yourself! Enjoy!'. On the right side, there is a 'Term of the Day' section with two entries: 'EN sick leave' and 'FR congé de maladie'. Below that is a section titled 'Learn About Best Practices in Multilingual Terminology' with a sub-heading 'Download a FREE publication' and the text 'Towards Consolidation of European Terminology Resources'. At the bottom of this section is a small image of a book cover.

EuroTermBank

- www.eurotermbank.eu
 - MS Word
 - memoQ
 - Microsoft multilingual terminology
 - IATE
 - Open Terminology Platform
 - sharing & exchange terminology in META-SHARE



- will be used in terminology services both for human & machines as users



ACCURAT & TTC

Analysis and Evaluation
of Comparable Corpora
for Under-Resourced Areas
of Machine Translation



Terminology Extraction
Translation Tools

Comparable Corpora



ACCURAT & TTC

- Comparable corpora
- Reference term lists and annotated texts
- Rule sets for term variant recognition and mapping
- Toolkit for multi-level alignment and information extraction from comparable
- Neo-classical multi-word term detection program
- TTC TermSuite

TaaS

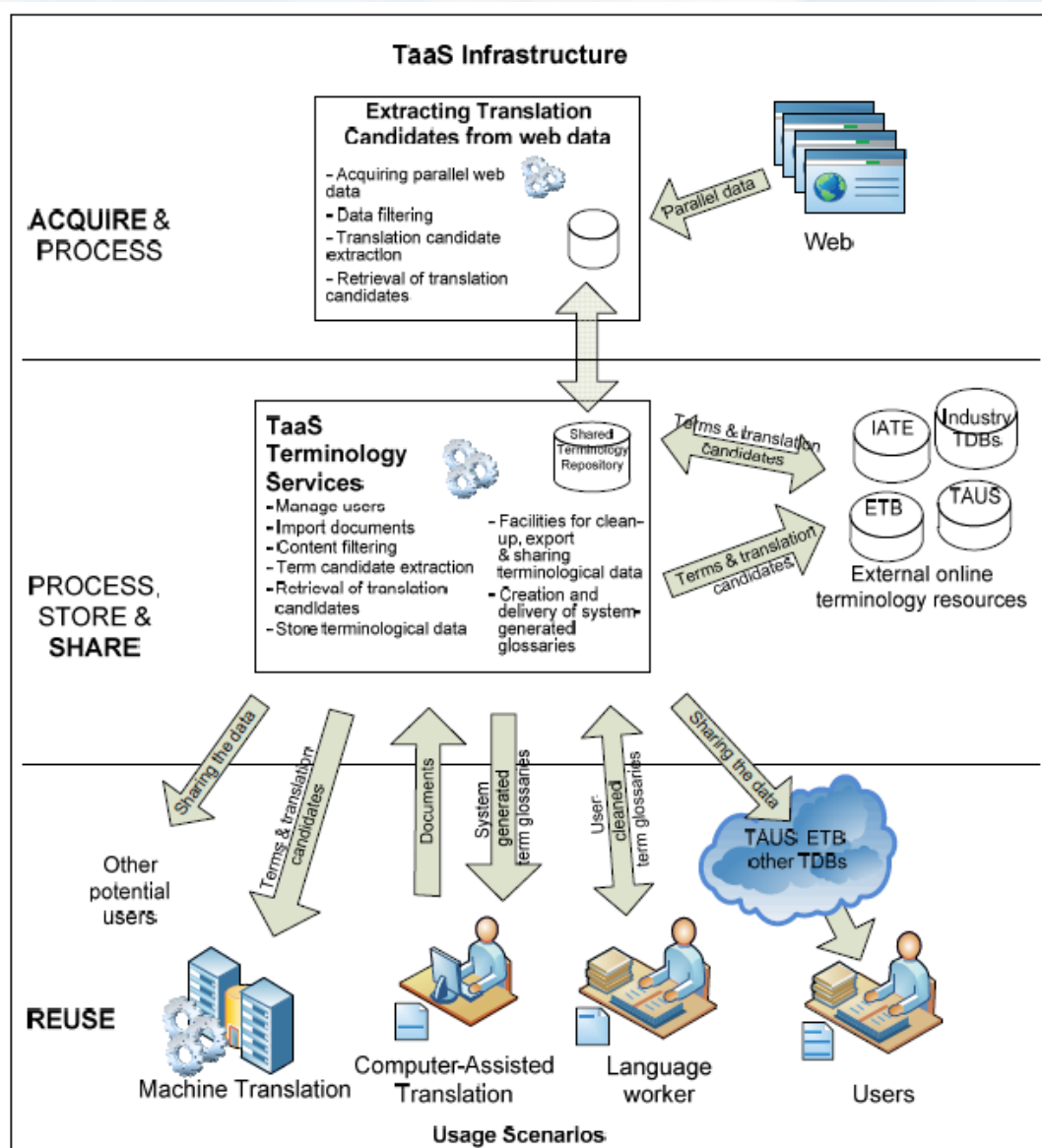
Terminology as a service

a cloud-based platform

for acquiring, cleaning up, sharing, and reusing
multilingual terminological data



TaaS basic services



LetsMT!

Build your own machine translation system!

With LetsMT you can easily build and run your own custom machine translation systems. Simply upload your own corpora and/or choose to use any of the publicly available corpora. Train your systems and use them for all your translation needs.

EASY

- ✓ Frees you from hardware and software infrastructure
- ✓ Store corpora and engines in one place
- ✓ Access to many free public texts and translation engines
- ✓ Instantly increase productivity with CAT plug-in for localization process and web browser widget

IS LETS MT FOR YOU?

- ✓ Localization & translation service providers
- ✓ Holders of linguistic resources
- ✓ Companies with a need to translate large amounts of information
- ✓ Those who don't want to trust their resources to public systems

THE BASICS



Translate

Translate now using available systems



Build

Build your own machine translation system



Store and Share

Upload, organize, store, and share your corpora

SMT adaptation use case

SMT system adaptation to narrow domain

- automotive manufacturing

We had:

- limited amount of in-domain parallel texts from a client
- no in-domain texts in the target language
- extracted terms from parallel texts
- additional comparable texts collected from the web
- bilingual in-domain terms tagged and mapped automatically in the collected texts

We got:

- **32%** increase in BLEU against a broad domain system

Terminology is on the cusp between
semantic and language technologies

Terminology bridges the three communities
LOD, MW & NLP

Terminology has the potential to vastly enhance
the degree of automation for LOD

Terminology facilitates the creation
of multilingual ontologies, taxonomies, etc.

Terminology helps to automate the creation
of multilingual & cross-lingual metadata

Thank you for your attention and time!

www.tilde.com

tatiana.gornostay@tilde.lv