ITS 2.0 Metadata & Machine Translation

Ankit Srivastava & **Declan Groves**

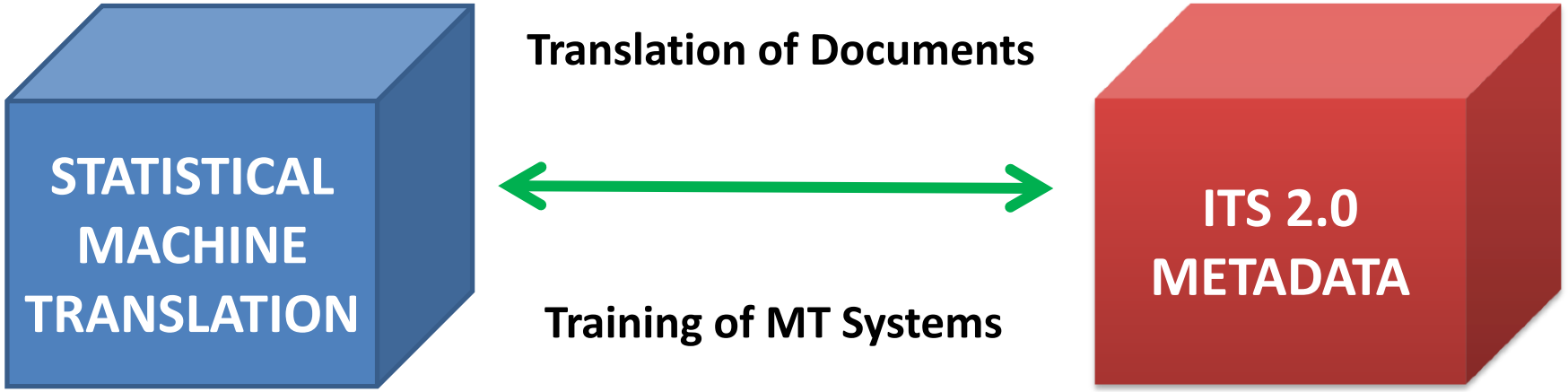
Centre for Next Generation Localisation (CNGL),

School of Computing, Dublin City University

ITS 2.0 Showcase @ Dublin

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Overview



WEB SERVICE DEMO

http://srv-cngl.computing.dcu.ie/mlwlt/

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Motivation

- Need for ITS2.0 and similar standards
  - MT inserted into the larger localization pipeline
  - Seamless integration and interoperability
- Purpose of the Web Services
  - Demonstration for MLW-LT (EU FP7 project)
  - Plans to release scripts as open source
    - Perspective of a Research University instead of a Company

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

TRANSLATION WEB SERVICE

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

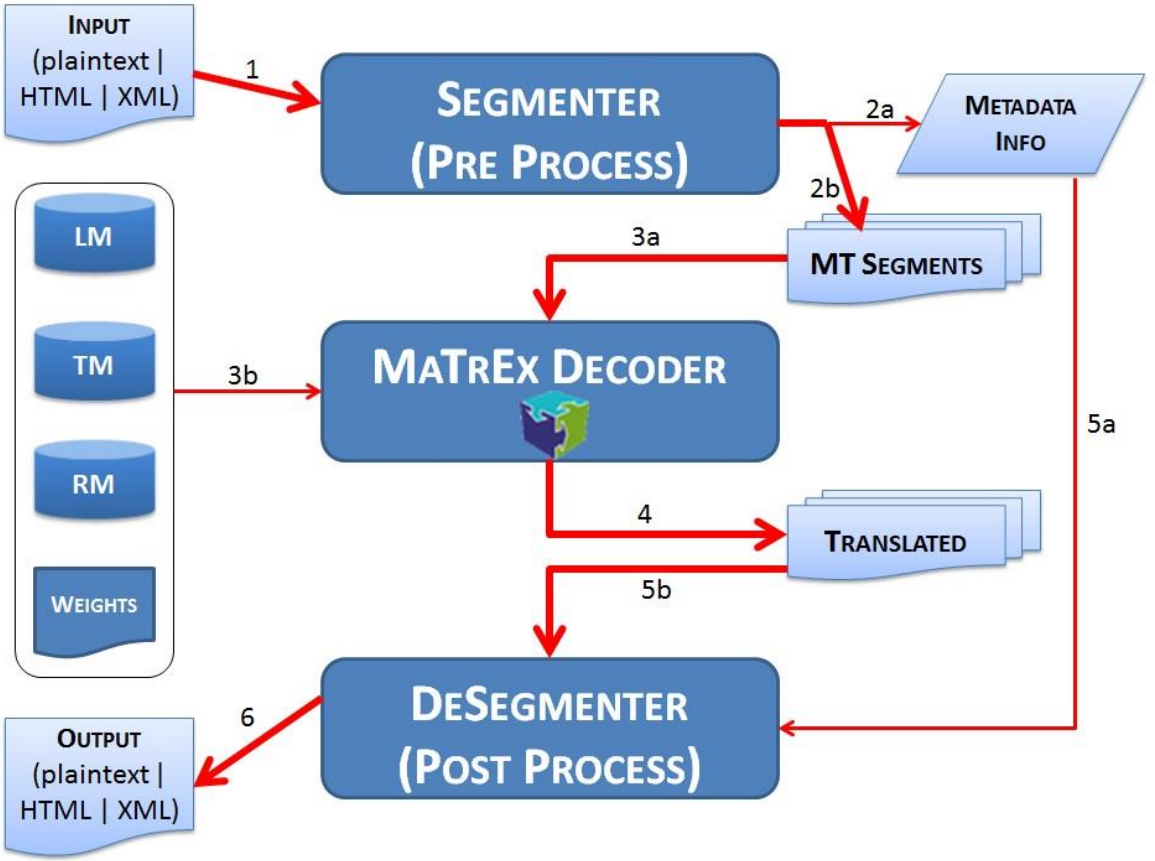
A Statistical Machine Translation (SMT) System developed in-house at DCU using the open-source Moses decoder

Segmenter takes as input ITS 2.0 tagged document, parses it, and generates segments to be translated [Pre Process]

Decoder translates the segments

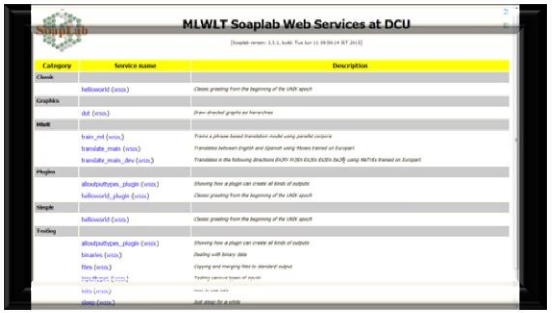
Desegmenter rebuilds the document by merging the translated segments with relevant ITS 2.0 metadata [Post Process]

A set of pre-processing and post-processing scripts wrapped around the MaTrEx Decoder enables us to translate ITS 2.0-tagged documents in HTML / XML / XLIFF



The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

DEMO of TRANSLATION WEB SERVICE



VIDEO

http://computing.dcu.ie/~asrivastava/docs/webservice_xlate_jun2013.wmv

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

MODIFICATIONS

- Translate Tagged Documents instead of Plain Text
- Input HTML / XLIFF DOCUMENT
- Extract Translatable Segments from Document
  - Document split into Segments
  - Special markups for <translate> <domain> <terminology>
  - Deal with special HTML constructs
    - Nested Tags, Overlapping Tags
- Decode using SMT Engine
- Reconstruct Document by re-inserting tags
  - Markup for MT Confidence
  - Take care of specific formatting like new-line, etc.
- Output Translated Document

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Example (Es to En)

SOURCE

Se informa de las principales novedades tributarias introducidas por la Ley 16/2012, de 27 de diciembre, por la que se adoptan diversas medidas tributarias dirigidas a la consolidación de las finanzas públicas y al impulso de la actividad económica.

EXPECTED OUTPUT

Information is provided on the main tax changes introduced by the Law 16/2012, of 27th of December, by which different tax measures are adopted in the aim of consolidating public finances and the impulse of the economic activity.

ACTUAL OUTPUT

Are reports of the main developments introducidas tax by law, 16 / 2012, 27, adopted by the various tax measures aimed at consolidating public finance, and Al boost económica activity.

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Example (Es to En)

SOURCE

Se informa de las principales novedades tributarias introducidas por la Ley 16/2012, de 27 de diciembre, por la que se adoptan diversas medidas tributarias dirigidas a la consolidación de las finanzas públicas y al impulso de la actividad económica.

EXPECTED OUTPUT

Information is provided on the main tax changes introduced by the Law 16/2012, of 27th of December, by which different tax measures are adopted in the aim of consolidating public finances and the impulse of the economic activity.

ACTUAL OUTPUT

Are reports of the main developments **introducidas** tax by law, **16 / 2012**, **27**, adopted by the various tax measures aimed at consolidating public **finance**, and Al boost **económica** activity.

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

BENEFITS: TRANSLATE

WITHOUT ITS

Are reports of the main developments introducidas tax by law, **16 / 2012**, 27, adopted by the various tax measures aimed at consolidating public finance, and Al boost económica activity.

WITH ITS

Are reports of the main developments introducidas tax by law, **16/2012**, 27, adopted by the various tax measures aimed at consolidating public finance, and Al boost económica activity.

…la Ley   16/2012, de 27 de diciembre, …

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

BENEFITS: DOMAIN

WITHOUT ITS

Are reports of the main developments introducidas tax by law, 16 / 2012, 27, adopted by the various tax measures aimed at consolidating public finance, and Al boost **económica** activity.

WITH ITS

Are reports of the main developments introducidas tax by law, 16 / 2012, 27, adopted by the various tax measures aimed at consolidating public finance, and Al boost **economic** activity.

Domain Rule for translating economic-specific terms
from ECON translation model

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

BENEFITS: TERMINOLOGY

WITHOUT ITS

Are reports of the main developments **introducidas** tax by law, 16 / 2012, **27**, adopted by the various tax measures aimed at consolidating public **finance**, and Al boost económica activity.

WITH ITS

Are reports of the main developments **introduced** tax by law, 16 / 2012, **27 December**, adopted by the various tax measures aimed at consolidating public **finances**, and Al boost económica activity.

TermInfo Pointer for terms which occur frequently in these texts on Taxation Law:
Introducidas = introduced ||| finanzas = finances ||| Month Names

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

BENEFITS: MT CONFIDENCE

- Of importance to successive stages in the localization cycle; Example Post Editors
- Generated by MT Engines
  - ITS enables marking specific segments with information about the accuracy and origin of translation

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

TRAINING WEB SERVICE

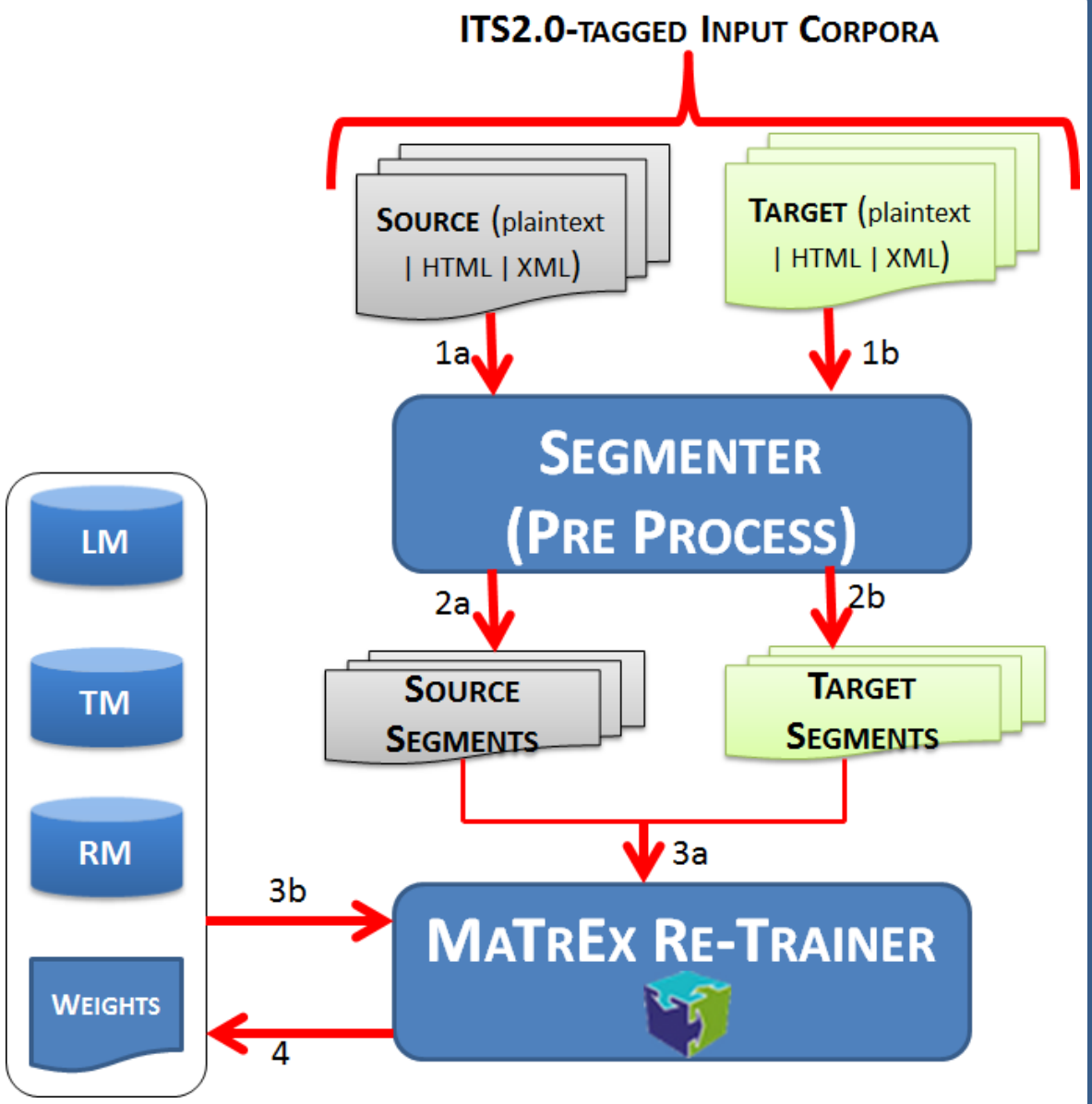
The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

MaTrEx

SMT components that could be potentially trained / retrained include a Language Model (LM), Translation Model (TM), Reordering Model (RM), and associated feature weights

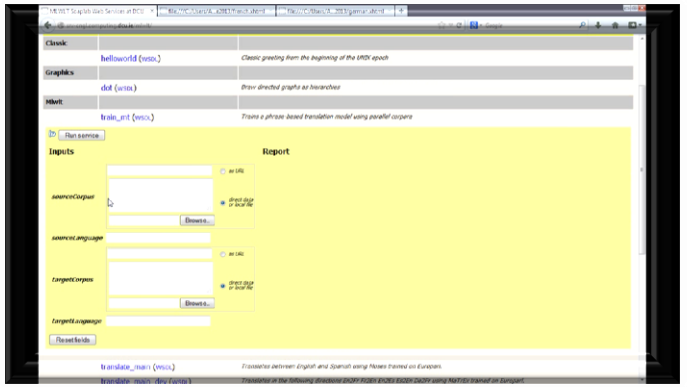
Segmenter takes as input ITS 2.0 tagged documents (sentence-aligned parallel content), parses it, and generates bilingual segments for training [Pre Process]

Retrainer processes bilingual segments to augment pre-existing SMT Models



The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

DEMO of TRAINING WEB SERVICE



VIDEO

http://computing.dcu.ie/~asrivastava/docs/webservice_train_jun2013.wmv

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Description

- Passive Training
  - Take as input source and target documents tagged with ITS 2.0
  - Extract translatable segments from both source and target
    - Reusing pre-processor scripts from Translation Service
    - **Remove all instances of any <…> from text**
  - Run MT Training Scripts
- Active Training

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Description (contd.)

- Passive Training
- Active Training
  - Take as input source and target documents tagged with ITS 2.0
  - Extract translatable segments from both source and target
    - Reusing pre-processors scripts from Translation Service
    - **Deal with specific tags like <translate> <domain>**
  - Run MT Training Scripts

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Training Experiments

- Motivation
  - Force the MT trainer to acknowledge terms which do not translate.
- Procedure
  - Take as input source and target documents tagged with ITS 2.0
  - Extract translatable segments from both source and target
    - Reusing pre-processors scripts from Translation Service
    - **Run Term Substitution Algorithm for <translate> elements**
      - **Replace with unique codes across the entire training data**
  - Run MT Training Scripts
    - **Revert unique codes to actual text**
  - Compare Results with Baseline
- Results
  - Inconclusive on Automatic MT Evaluation
  - Needs further investigation

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Conclusions

- ITS data categories currently implemented
  - Domain, Language Information, Locale Filter, MT Confidence, Provenance, Terminology, Translate

ITS 2.0 Data Category	Usage / Benefits to the System
Domain	Enables seamless application of domain-tuned MT engines
Language Info	Enables easy identification of source / target language
Locale Filter	Enables locale-specific translation operations
MT Confidence	Enables scores to be displayed accurately and automatically to PEs
Provenance	Enables localisation workflow managers to compare performance
Terminology	Enables terms to be identified for translations to be enforced in MT
Translate	Ensures identification of text fragments that need to be preserved

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Conclusions Contd.

- MT is just a small part of a larger workflow
  - In order to incorporate in-house MT system in commercial translation / localization projects (larger pipeline), compliance with a standard is essential
  - Benefits are interoperability, customizability (terminology, domain, etc.)
- Training with information stored in metadata is possible
  - Needs further investigation
- Demonstrate pre / post wrapper scripts are sufficient to adapt a pre-existing MT system to the ITS 2.0 standard

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

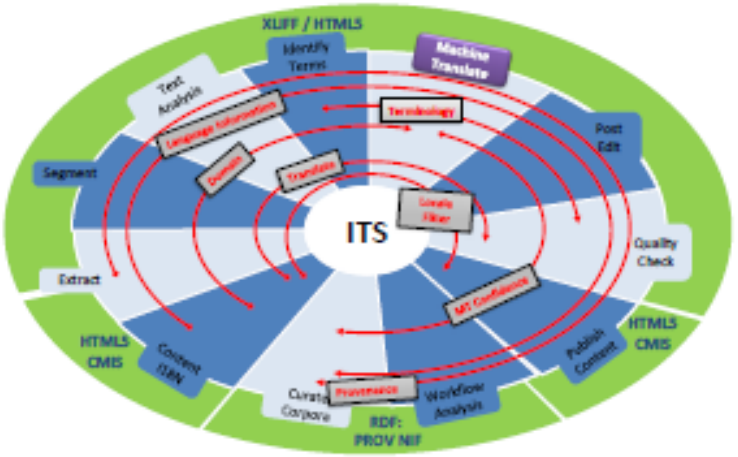
Acknowledgements

Spanish – English Data

German – French Data

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Thank You!



The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

EXPERIMENTS: translate tags

- BASELINE
- BASELINE + ITS (passive)
- BASELINE + ITS (active)

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

ITS 2.0-*enabled* Statistical Machine Translation and Training Web Services

Ankit Srivastava & **Declan Groves**

Centre for Next Generation Localisation (CNGL),

School of Computing, Dublin City University

ITS 2.0 Showcase @ Dublin

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.