PRINCIPLE

ELRC Conference 10th March

Jane Dunne – Dublin City University (DCU)

Co-financed by the Connecting Europe Facility of

PRINCIPLE Introduction

- Overview
- Use Case Analysis
- Data Requirements and Preparation
- Development of MT systems
- Evaluation of MT systems
- Work remaining on project

PRINCIPLE Overview of Project

Consortium Members: *Dublin City University (Project Coordinator), University of Iceland, Faculty of Humanities and Social Sciences, University of Zagreb, National Library of Norway, Iconic Translation Machines Ltd.*

- 2-year Connecting Europe Facility (CEF) project
- Started in September 2019
- Focus on **data collection** to improve translation quality in the DSIs of *eJustice* and *eProcurement*

@principle_proj

PRINCIPLE Project

www.principleproject.eu

Overview of Project

Goal - Identify, collect and process high-quality Language Resources (LRs) for:

- Croatian
- Icelandic
- Irish
- Norwegian (Bokmål and Nynorsk)

PRINCIPLE Overview of Project

PRINCIPLE will identify high-quality curated LRs via domain-specific MT engines

MT engines will be offered to the early adopter partners in Croatia, Iceland, Ireland and Norway for the duration of the project

The MT systems built will be evaluated to demonstrate the benefits of the project

Data **deemed to be of high-quality** will be uploaded to ELRC-SHARE repository to improve the eTranslation engines

Use-case analysis

Two use-case scenarios were identified by the consortium

1. Data contributors

   Public and government bodies & industry partners (in each country) that are aligned with the **specific** domains

2. ‘Early adopters’

   ‘Early adopters’ would be provided with **domain-specific MT engines** for the **duration of the project**

PRINCIPLE Use-case analysis

Data contributors in all 4 countries completed questionnaires to gauge:

- Translation process – needs, demands, workflows
- Type of LRs available – formats, quality and quantity

Questionnaire for Data Contributors

About Data Contributors	
Organization	
Name:	
Address:	
URL:	
Contact person	
First name:	
Last name:	
Email:	
Phone number:	
Address:	
"Early adopter": (select one)	• Yes • No
About the Translation Process	
Is translation part of your workflow? If so, describe the use-case(s).	
In what file format is the data you receive that needs to be translated? E.g. plain text files, Microsoft Word format, PDF files, TMX, TBX, XLIFF, etc.	

Data requirements and preparation

Consortium agreement

1) LRs require language identification
2) Acceptable file formats are the following: .TXT, .DOC(X), TMX, TBX, XLIFF, .PDF, .XLSX
3) Parallel corpora & monolingual corpora
4) Sentence aligned texts preferred (automatic/manual alignment carried out otherwise)
5) All data pre-processing has to be documented

PRINCIPLE Development of MT Systems

Iconic Translation Machines completed a full review and quality check on ELRC Resources

ELRC-SHARE Data used for 1st Baseline Engines*

Language	No. of TUs
Irish	588,663
Croatian	3,337,608
Icelandic	702,139
Norwegian (Bokmål)	1,140,351
Norwegian (Nynorsk)†	-

[*After Iconic cleaning/filtering of bi-lingual corpora]
[† a lack of public data for Nynorsk meant that it was not possible to train a Nynorsk engine]

Internal MT Evaluation Steps

- Using **3,000 sentence test set** automatic metrics are computed:
  - BLEU
  - METEOR
  - TER
  - chrF
- Same **3,000 sentence test set** is translated through **eTranslation** and the public **Google** and **Microsoft** interfaces
- Automatic metrics for eTranslation, Google and Microsoft are computed and **compared to Iconic baselines**

PRINCIPLE **Phase 1** Early Adopters (MT engines created to date)

Data provider	Country	Domain
National University of Ireland Galway (NUIG)	Ireland	eProcurement
CIKLOPEA D.O.O	Croatia	eProcurement
Icelandic Ministry of Foreign Affairs	Iceland	eJustice / eProcurement
Standards Norway	Norway	eProcurement
Norwegian Ministry of Foreign Affairs	Norway	eJustice

NORWEGIAN MINISTRY OF FOREIGN AFFAIRS

CIKLOPEA

OÉ Gaillimh NUI Galway

UTANRÍKISRÁÐUNEYTIÐ Ministry for Foreign Affairs Iceland

standards norway

PRINCIPLE Evaluation of MT Systems

- **MT evaluation protocol** prepared and agreed
- Instructions and guidelines to prepare test sets (500 sentences) **shared with EAs**
- ‘Menu’ of options for automatic and human MT evaluation shared with EAs
- **Agree** on specific evaluation models (**especially human methods**) relevant to each EA use-case

PRINCIPLE **Phase 2** Early Adopters (MT engines in development)

Data provider	Country	Domain
Rannóg an Aistriúcháin	Ireland	eJustice
Foras na Gaeilge	Ireland	eProcurement
CIKLOPEA D.O.O	Croatia	eHealth
Ministry of Foreign and European Affairs	Croatia	eJustice
Icelandic Standards	Iceland	eJustice / eProcurement
Icelandic Meteorology Office	Iceland	Other (Meteorology)

ÍSLENSKIR STAÐLAR
STAÐLARÁÐ ÍSLANDS
CIKLOPEA
Foras na Gaeilge
Tithe an Oireachtais
Houses of the Oireachtas
Veðurstofa Íslands
REPUBLIC OF CROATIA
Ministry of Foreign and European Affairs

PRINCIPLE Final 6 months - ToDo

- Confirm **Early Adopters for Phase 2**
- Develop **MT systems** for Phase 2 EAs
- **Evaluate** Phase 2 MT systems by April 30th
- **Deploy** MT engines to all Phase 2 EAs
- **Validate "high quality" data** and **upload to the ELRC-SHARE platform**
- **Promotional workshops**

PRINCIPLE

Go raibh maith agaibh!

Þakka þér fyrir

Hvala vam

Takk skal du ha

Thank you!