PRINCIPLE

ELRC Workshop 24th June

Jane Dunne – Dublin City University (DCU)

Co-financed by the Connecting Europe Facility of the European Union

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

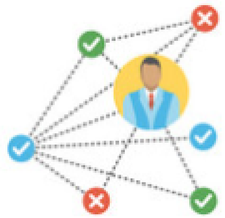
PRINCIPLE Overview of Project

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

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

PRINCIPLE Overview of Project

PRINCIPLE has identified high-quality curated LRs via domain-specific MT engines



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

The MT systems built were evaluated to demonstrate the benefits of the project

European Language Resource Coordination
Connecting Europe Facility

Data **deemed to be of high-quality** continues to be uploaded to ELRC-SHARE repository to improve the eTranslation engines until the end of the project (August 2021)

PRINCIPLE 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.	

PRINCIPLE Data requirements and preparation

The PRINCIPLE consortium agreed on the following:

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)

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]

PRINCIPLE Early Adopters (MT engines created to date)

Phase 1

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

Phase 2

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)

REPUBLIC OF CROATIA
Ministry of Foreign and European Affairs

CIKLOPEA

ÍSLENSKIR STAÐLAR
STAÐLARÁÐ ÍSLANDS

Veðurstofa Íslands

Tithe an Oireachtais
Houses of the Oireachtas

NORWEGIAN MINISTRY OF FOREIGN AFFAIRS

OÉ Gaillimh
NUI Galway

UTANRÍKISRÁÐUNEYTIÐ
Ministry for Foreign Affairs Iceland

standards norway

Foras na Gaeilge

PRINCIPLE Evaluation of MT Systems (DCU)

- **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 Evaluation of MT Systems (DCU)

- Automatice evaluation - Iconic EA Engines **outperformed** Google Translate, Microsoft Translator and eTranslation in **majority of cases**
- Human evaluation - high number of fluent and adequate translations - **translating with MT considered faster than translating from scratch**
- Results of **both the human and the automatic evaluations** confirm **quality of the Iconic MT systems**, but also indirectly **validate the quality** of the collected data by the partners.

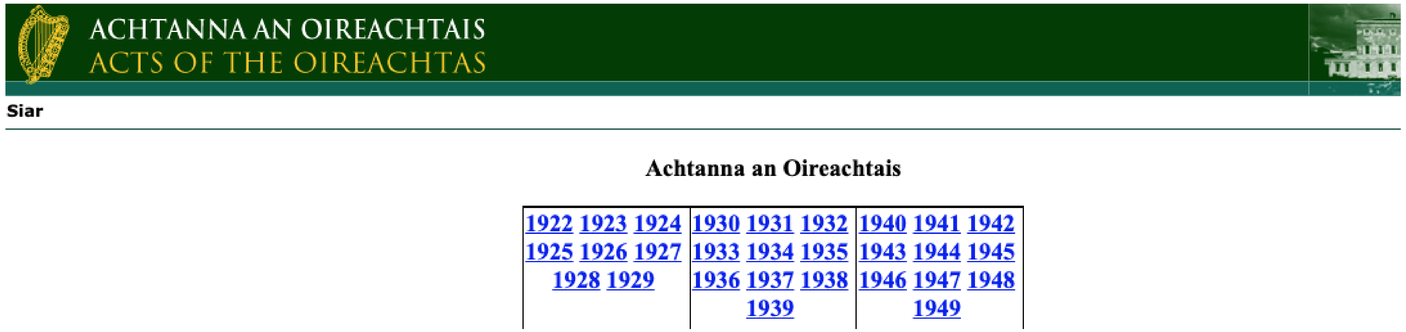
PRINCIPLE Sources of Irish language data collected

TMX files from translators using CAT tools

TMX files from LSPs (using CAT tools) employed by EAs

PDFs of Acts collected from websites e.g. oireachtas.ie

Website scraping e.g. achtanna.ie


ACHTANNA AN OIREACHTAIS
ACTS OF THE OIREACHTAS

Siar

Achtanna an Oireachtais

1922 1923 1924	1930 1931 1932	1940 1941 1942
1925 1926 1927	1933 1934 1935	1943 1944 1945
1928 1929	1936 1937 1938	1946 1947 1948
	1939	1949

Tithe an Oireachtais Houses of the Oireachtas

An Roinn Dlí agus Cirt Department of Justice

Foras na Gaeilge

OÉ Gaillimh NUI Galway

An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meán
Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media

PRINCIPLE Overview of Irish language data collected

Data provider	Domain	Dataset	Amount of data (translation units)
National University of Ireland Galway (NUIG)	General/eProcurement	All	17,949
Foras na Gaeilge	General/eProcurement	All	60,443
Department of Justice	eJustice	Secondary Legislation	35,898
DCHG	General	All	64,694
Rannóg an Aistriúcháin	eJustice	Primary Legislation	463,530
		Secondary Legislation	28,487
		Ancillary	37,729
	General	Annual Reports	7,512
Total			**716,242**

Tithe an Oireachtais
Houses of the Oireachtas

An Roinn Dlí agus Cirt
Department of Justice

Foras na Gaeilge

OÉ Gaillimh
NUI Galway

An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meán
Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media

PRINCIPLE Concluding work until end of project (Aug 2021)

- Deploy MT engines to all Phase 2 EAs
- Validate “high quality” data and upload to the ELRC-SHARE platform

- Remaining workshops promoting the project and the benefits of language technology taking place

PRINCIPLE

Go raibh maith agaibh!

Þakka þér fyrir

Hvala vam

Takk skal du ha

Thank you!