Building MT systems in low resourced EU languages for Public Sector users in Croatia, Iceland, Ireland and Norway.

Páraic Sheridan

MT Summit: August 2021

PRINCIPLE

The work presented here is co-financed by the Connecting Europe Facility of the European Union

RWS

[Re-]Introducing Language Weaver

The first mile

1949
Warren Weaver
Translations Memorandum

2002
Language Weaver

2012
Iconic Translations

TODAY

The last mile

Language Weaver. The last mile in machine translation.

RWS

Introducing The PRINCIPLE Project

- A 2-year project funded by the Connecting Europe Facility (CEF)
- Focused on collecting data to improve translation quality in the EU Digital Services Infrastructures (DSIs) for prioritised low-resourced EU languages.
- The main aim of the project is to identify, collect and process high-quality Language Resources (LRs) for the following under-resourced European languages:
  - Croatian
  - Icelandic
  - Irish
  - Norwegian (Bokmål and Nynorsk)

Project Consortium:

PRINCIPLE: The Role of Machine Translation

By building state-of-the-art Neural MT models with data collected in the PRINCIPLE project, two key objectives can be accomplished:

✓ Benchmarking and evaluation of MT systems built using project data attests to the quality of data collected and its value for MT systems developed in Europe.

Granting free access and use of MT systems to Public Sector bodies during the course of the project provides an incentive for contributions of language data.

- Public sector bodies who participate in this incentive are labelled '**Early Adopters**' in the PRINCIPLE project.

PRINCIPLE

RWS

What Data Already Existed for These Languages?

Iconic completed a full search/download of existing resources from ELRC-Share*.

A quality review was conducted by PRINCIPLE project partners.

Language	# Resources	# Translation Units
Irish	41	901,421
Croatian	36	3,891,799
Icelandic	17	801,283
Norwegian	47	1,964,961
Norwegian (Nynorsk)	4	6,358

PRINCIPLE

* https://elrc-share.eu/

RWS

What Data Already Existed for These Languages?

Iconic completed a full search/download of existing resources from ELRC-Share.

A quality review was conducted by PRINCIPLE project partners.

Data was then cleaned/filtered for MT Baseline system development.

Language	# Resources	# Translation Units	#TU used in MT Baseline
Irish	41	901,421	588,663
Croatian	36	3,891,799	3,337,608
Icelandic	17	801,283	702,139
Norwegian	47	1,964,961	1,140,351
Norwegian (Nynorsk)	4	6,358	-

PRINCIPLE

* https://elrc-share.eu/

RWS

The PRINCIPLE Project then proceeded in Two Phases:

1

Data Provider	Country
National University of Ireland Galway (NUIG)	Ireland
CIKLOPEA D.O.O	Croatia
Icelandic Ministry of Foreign Affairs	Iceland
Standards Norway	Norway
Norwegian Ministry of Foreign Affairs	Norway

standards norway

UTANRÍKISRÁÐUNEYTIÐ
Ministry for Foreign Affairs Iceland

OÉ Gaillimh
NUI Galway

NORWEGIAN MINISTRY
OF FOREIGN AFFAIRS

CIKLOPEA

2

Data Provider	Country
Rannóg an Aistriúcháin	Ireland
Foras na Gaeilge	Ireland
CIKLOPEA D.O.O	Croatia
Ministry of Foreign and European Affairs	Croatia
Icelandic Standards	Iceland
Icelandic Met Office	Iceland

Tithe an Oireachtais
Houses of the Oireachtas

REPUBLIC OF CROATIA
Ministry of Foreign and European Affairs

Veðurstofa Íslands

Íslenskir staðlar

CIKLOPEA

Foras na Gaeilge

PRINCIPLE

RWS

Language Resources in PRINCIPLE.

Language Weaver. The last mile in machine translation.

Language Resources collected in PRINCIPLE - Croatian

Dataset	TUs Collected	Data used in MT
EN>HR Baseline	3,891,799	3,708,493
MVEP Data	115,667	100,649
Other Data Providers	22,703	

REPUBLIC OF CROATIA
Ministry of Foreign and European Affairs

Dataset	TUs Collected	Data used in MT
HR>EN Baseline	3,891,799	3,708,493
Ciklopea Data (eProcurement)	36,634	47,135
Other Data Providers	22,703	

CIKLOPEA
eProcurement

Dataset	TUs Collected	Data used in MT
EN>HR Baseline	3,891,799	3,708,493
Ciklopea Data (eHealth)	76,108	72,455

CIKLOPEA
eHealth

PRINCIPLE

RWS

Language Resources collected in PRINCIPLE - Irish

Dataset	TUs Collected	Data used in MT
EN>GA Baseline	901,421	588,663
Foras na Gaeilge	60,443	54,141
Rannóg an Aistriúcháin	387,480	353,485
Dept. Culture… & Gaeltacht	64,694	58,057

Foras na Gaeilge

Dataset	TUs Collected	Data used in MT
EN>GA Baseline	901,421	588,663
Rannóg an Aistriúcháin	387,480	353,485
Dept. of Justice	35,898	28,639
Dept. Culture… & Gaeltacht	64,694	58,057

Tithe an Oireachtais
Houses of the Oireachtas

An Roinn Dlí agus Cirt
Department of Justice

PRINCIPLE

RWS

Language Resources collected in PRINCIPLE - Icelandic

Dataset [EN<>IS]	TUs Collected	Data used in MT
Ministry of Foreign Affairs Data	1,097,352	821,243

Note that the Icelandic Ministry of Foreign Affairs stipulated only their data to be used, no baseline/other data.

UTANRÍKISRÁÐUNEYTIÐ
Ministry for Foreign Affairs Iceland

Dataset	TUs Collected	Data used in MT
IS>EN Baseline	801,283	702,139
Icelandic Met Office Data	214,242	188,700

Veðurstofa Íslands

Icelandic Met Office

Dataset	TUs Collected	Data used in MT
EN>IS Baseline	801,283	702,139
Standards Iceland Data	16,590	16,423

Íslenskir staðlar

PRINCIPLE

RWS

Language Resources collected in PRINCIPLE – Norwegian [Bokmål]

Dataset [EN>NO]	TUs Collected	Data used in MT
Norwegian Ministry Foreign Affairs	1,757,609	1,616,568

Note that the Norwegian Ministry of Foreign Affairs stipulated only their data to be used, no baseline/other data.

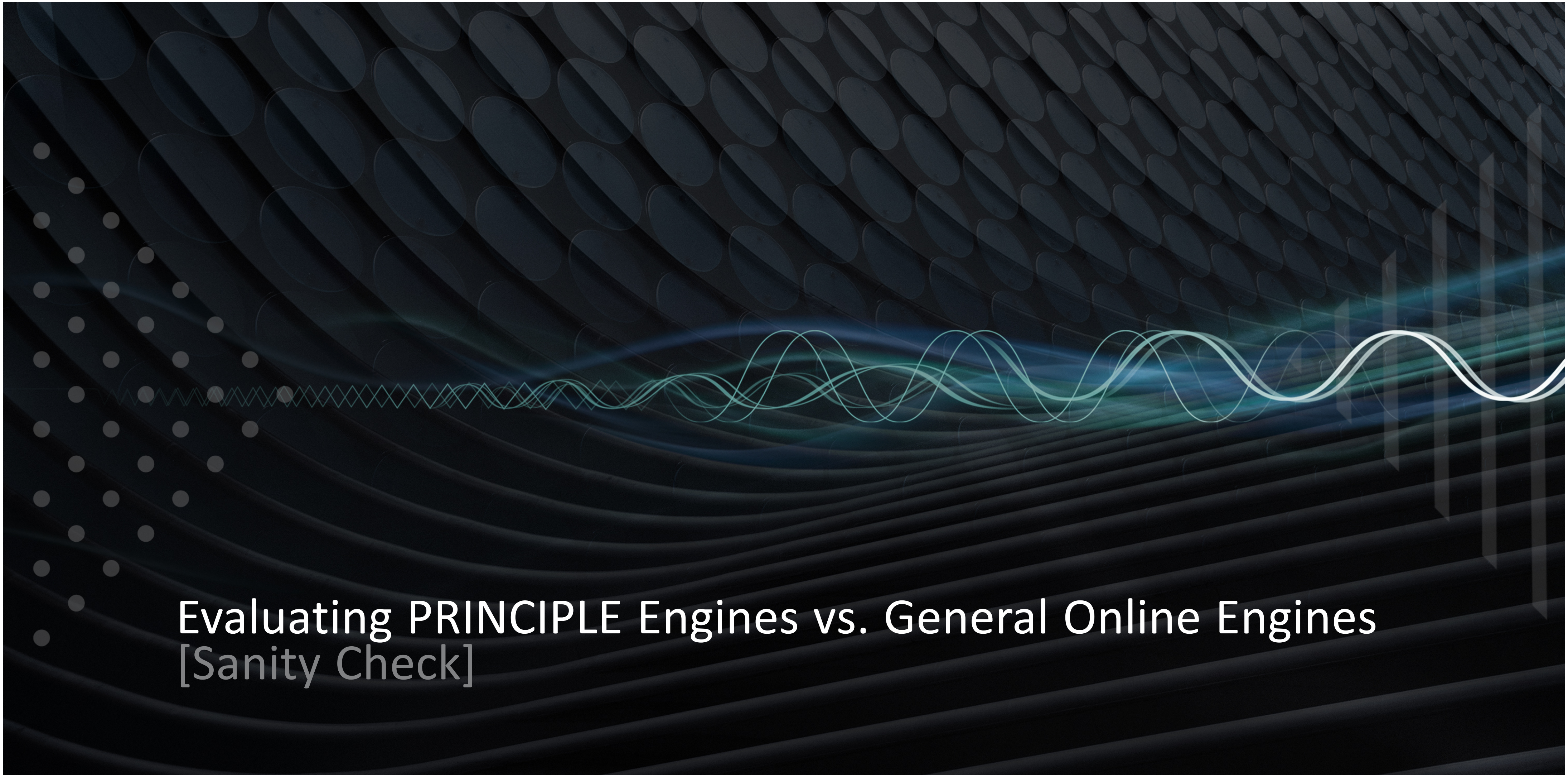
NORWEGIAN MINISTRY OF FOREIGN AFFAIRS

Dataset	TUs Collected	Data used in MT
EN>NO Baseline	1,964,961	1,140,351
Standards Norway Data	132,360	77,664

standards norway

PRINCIPLE

RWS



Evaluating PRINCIPLE Engines vs. General Online Engines

[Sanity Check]

Language Weaver. The last mile in machine translation.

An Overview of Automatic MT Evaluation in PRINCIPLE

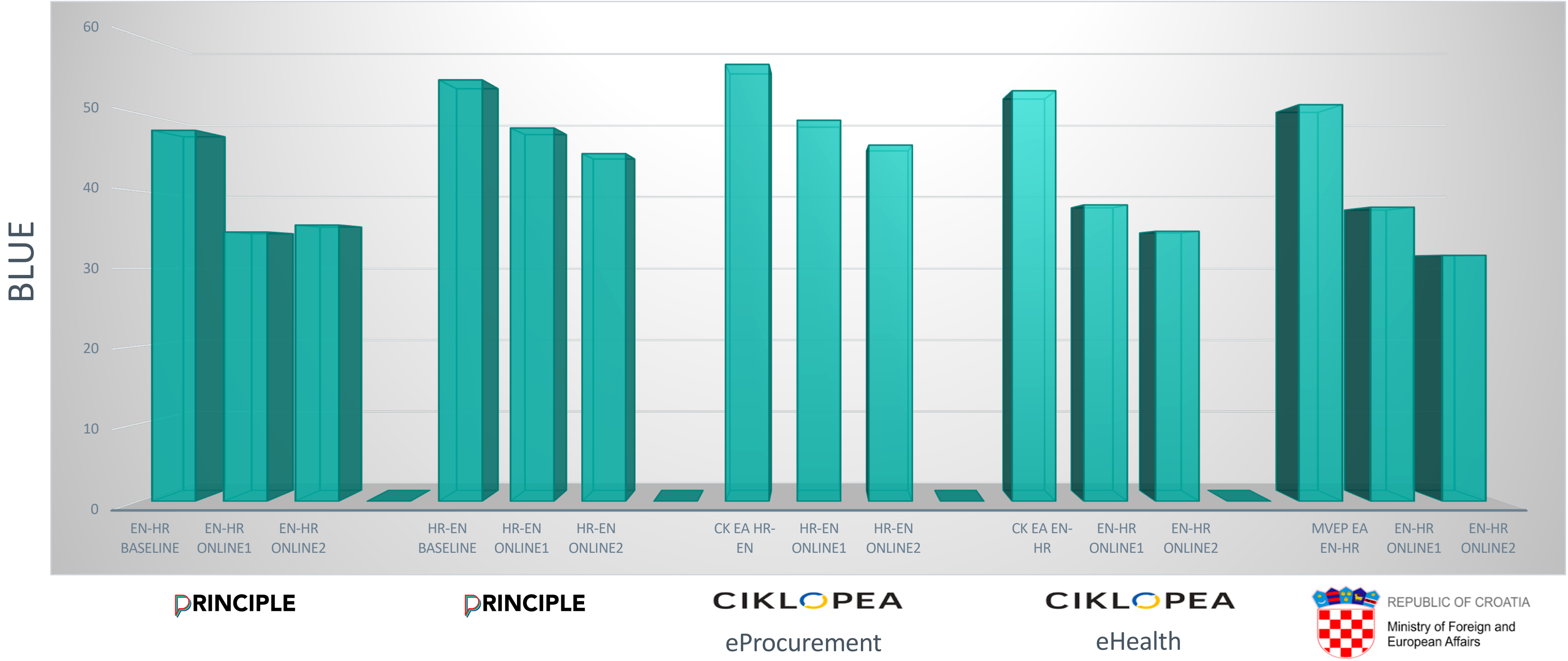
For every MT model developed by Iconic, a sanity-check evaluation was conducted against freely available online MT Engines.

A test set of 2,000 segments is generally held out as a test from data provided by customers. In some cases with PRINCIPLE Early Adopters, where limited data was provided, a test set of 1,000 segments or 1,500 segments was used.

Test segments are run through multiple MT engines for comparison, with a range of metrics computed [SacreBLEU, TER, METEOR, chrF].

- Each data set (bar triplets) represents the evaluation on a held-out test set for that model, either a baseline model for the language (PRINCIPLE), or a model with Early Adopter data.

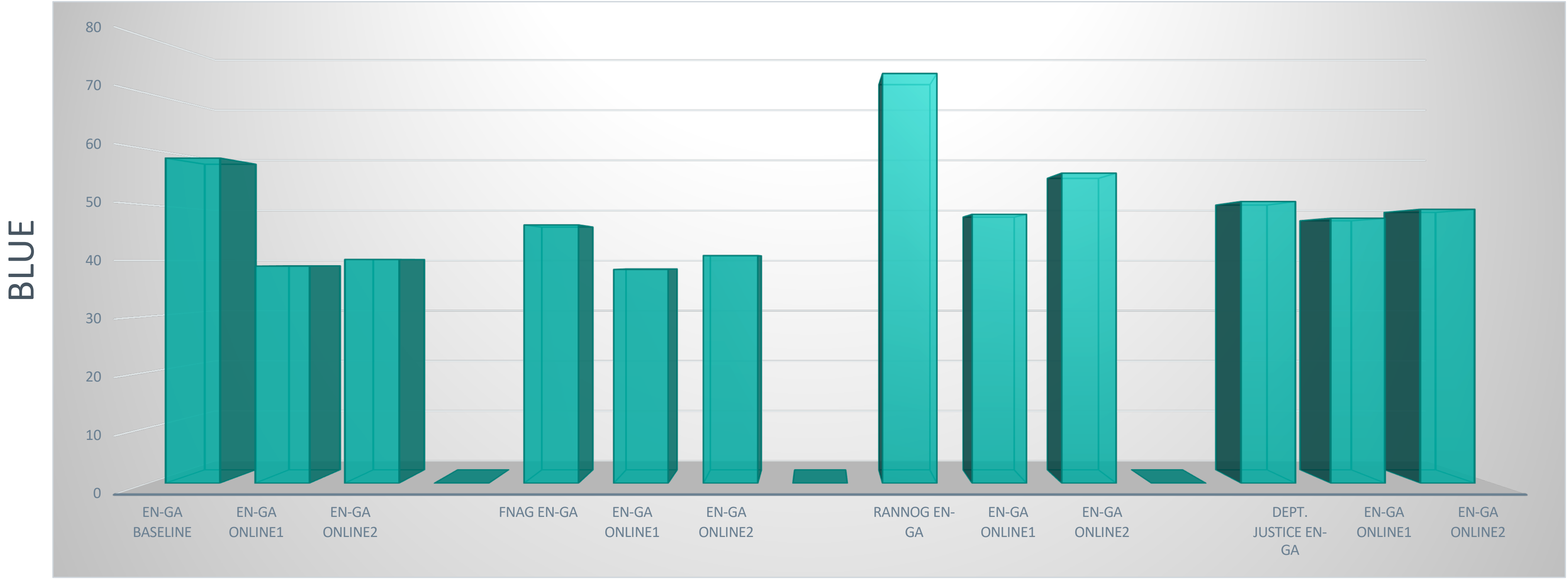
Comparing PRINCIPLE Engines to Online MT - Croatian



PRINCIPLE

RWS

Comparing PRINCIPLE Engines to Online MT – Irish



PRINCIPLE

Foras na Gaeilge

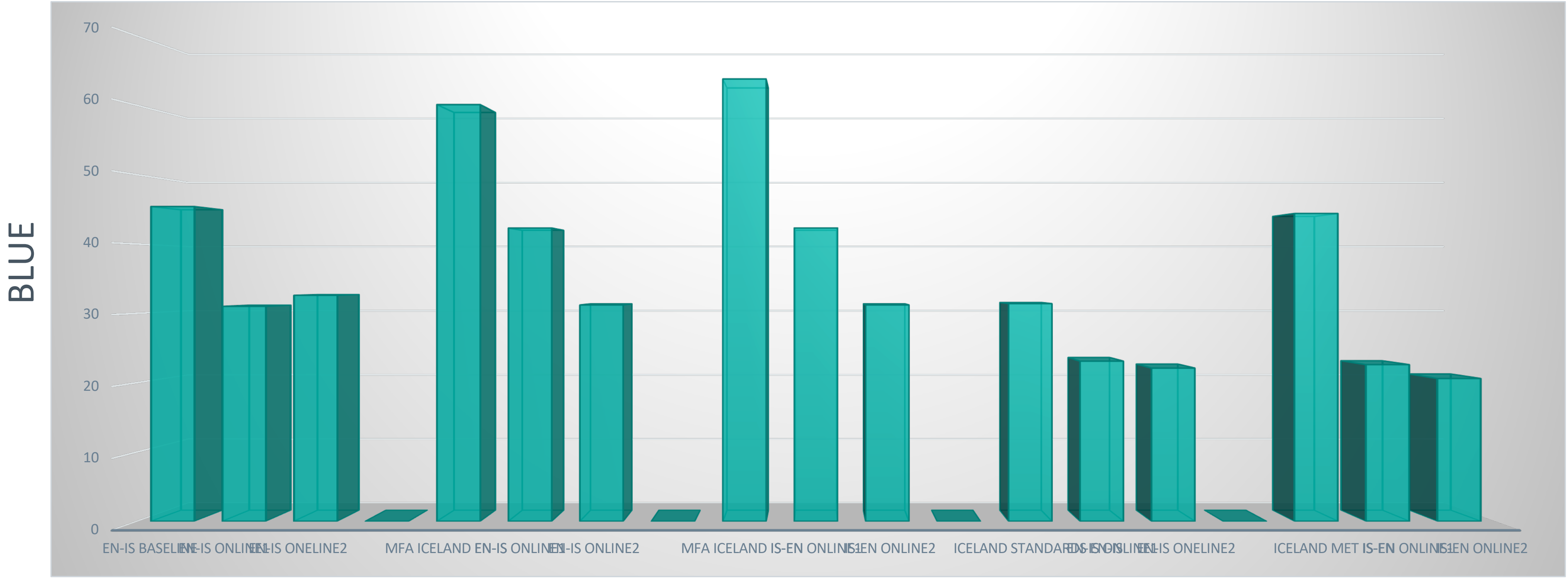
Tithe an Oireachtais Houses of the Oireachtas

An Roinn Dlí agus Cirt Department of Justice

PRINCIPLE

RWS

Comparing PRINCIPLE Engines to Online MT – Icelandic



PRINCIPLE

UTANRÍKISRÁÐUNEYTIÐ
Ministry for Foreign Affairs Iceland

UTANRÍKISRÁÐUNEYTIÐ
Ministry for Foreign Affairs Iceland

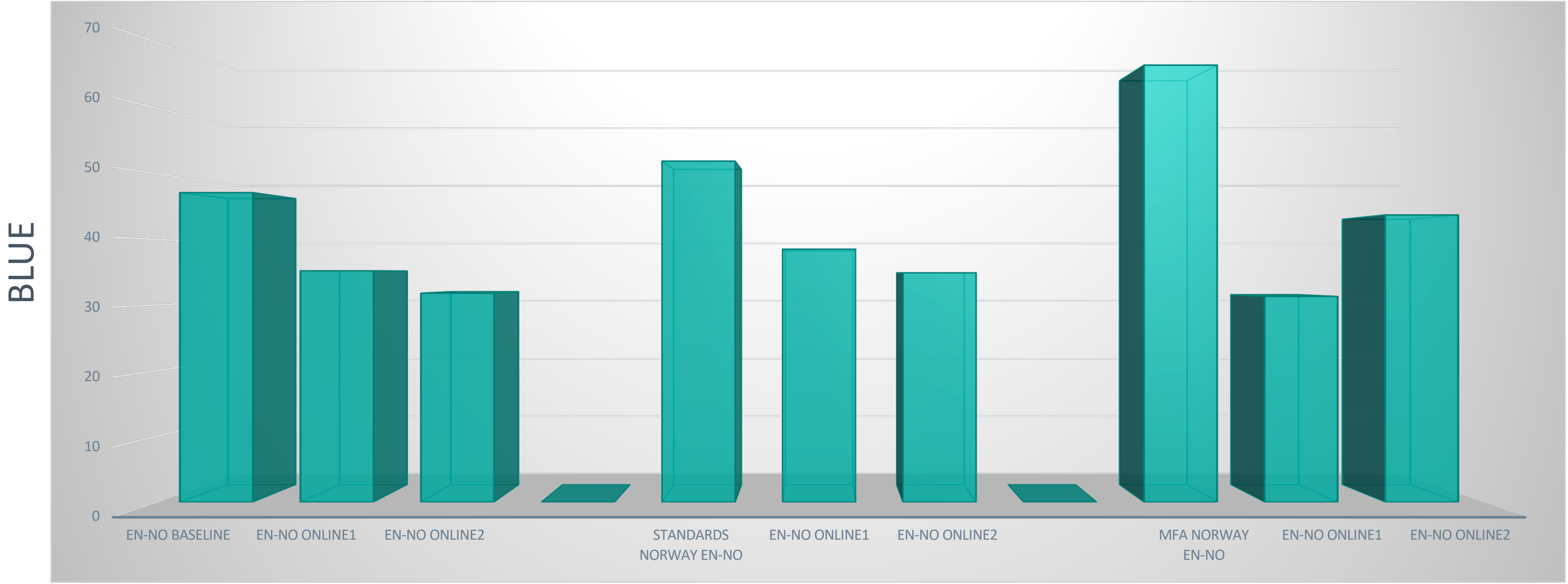
Íslenskir staðlar

Icelandic Met Office

PRINCIPLE

RWS

Comparing PRINCIPLE Engines to Online MT – Norwegian



PRINCIPLE

standards norway

NORWEGIAN MINISTRY OF FOREIGN AFFAIRS

PRINCIPLE

RWS

Sample User Evaluations

Language Weaver. The last mile in machine translation.

RWS

An Overview of User MT Evaluation in PRINCIPLE

Each PRINCIPLE ’Early Adopter’ was invited to develop a test set to be used by DCU (Evaluation co-ordinator) to help evaluate MT both using automatic and manual means.

A test set was requested of 500 segment pairs that

- Had not already been provided to train the MT systems.
- Were representative of the texts intended to be translated with the MT system.
- The reference translation in the target language should not be obtained via MT/Post-edit.
- Did not contain any confidential material.

Early Adopters were offered a range of human evaluation protocols from which they could choose, depending on their preference and available resources.

- Comparative ranking, adequacy & fluency, direct assessment, comprehension, post-editing, or MT error analysis

PRINCIPLE

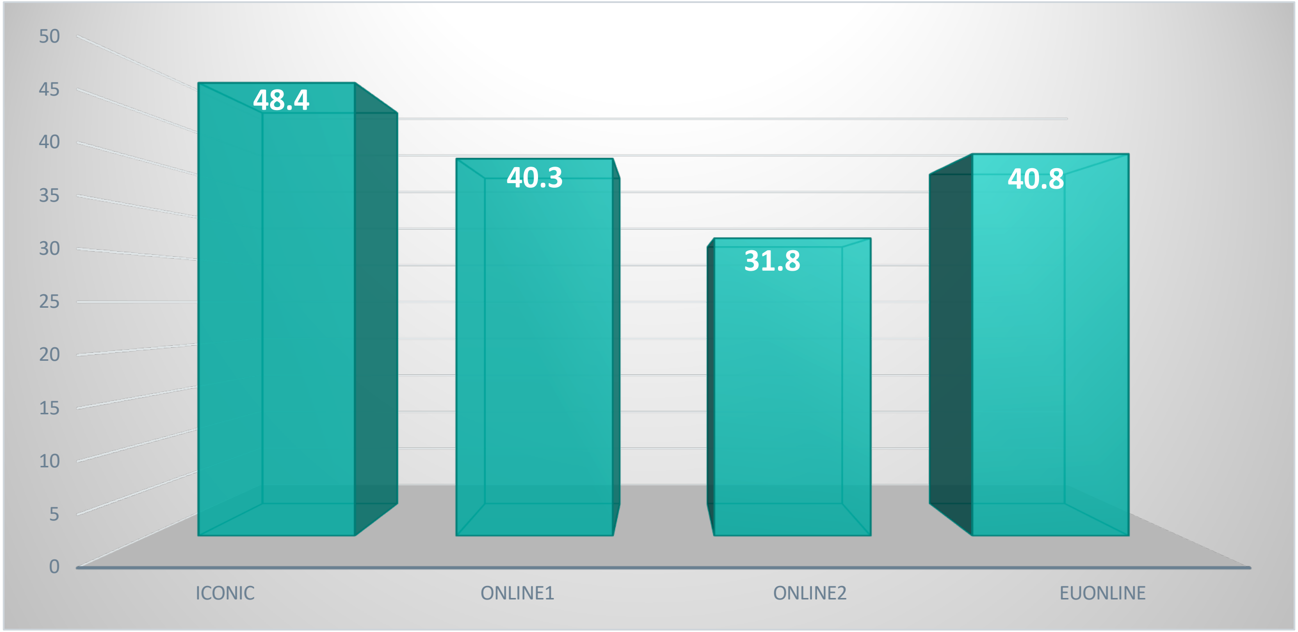
RWS

Comparison of MT Engines at Norwegian MFA [EN-NO]

NORWEGIAN MINISTRY
OF FOREIGN AFFAIRS

A 500-segment Test Set was created by MFA Norway, separate from all training data.

An automatic evaluation was conducted independently by DCU of four MT engines.



BLEU scores of four engines on a 500-segment test set provided by MFA Norway.

PRINCIPLE

RWS

Comparison of MT Engines at Norwegian MFA [EN-NO]

NORWEGIAN MINISTRY
OF FOREIGN AFFAIRS

A direct comparison of two engines was conducted by three evaluators at MFA Norway across the 500-segment test set (one evaluator completed only half of the test set).

For 70% of segments, Iconic's MFA engine was equal to or better than the comparator.

	Evaluator 1 (500 Segments)		Evaluator 2 (500 Segments)		Evaluator 3 (250 Segments)		Total (1,250 Judgements)	
Iconic Best	229	45.8%	260	52.0%	94	37.6%	**583**	**46.6%**
Online Best	138	27.6%	127	25.4%	68	27.2%	**333**	**26.6%**
Equally Good	118	23.6%	84	16.8%	86	34.4%	**288**	**23.0%**
Equally Poor	14	2.8%	29	5.8%	1	0.4%	**44**	**3.5%**
Not Assigned	1	0.2%	0	0.0%	1	0.4%	**2**	**0.1%**
Total	500	100%	500	100%	250	100%	**1,250**	**99.8%**

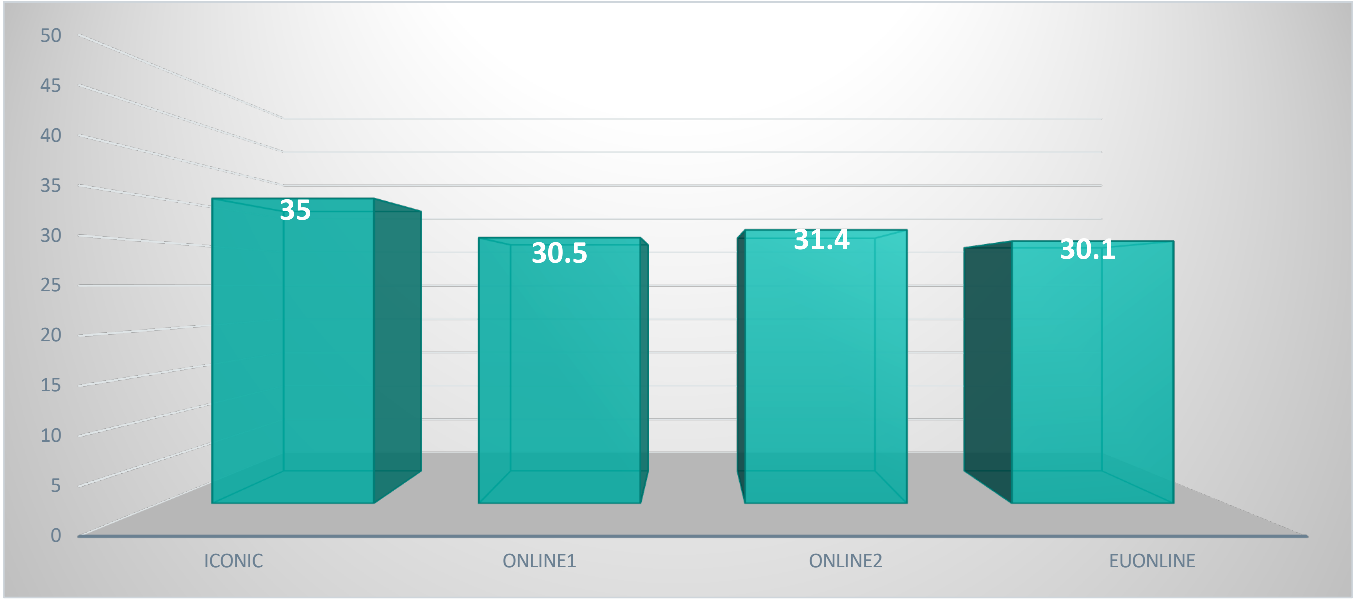
PRINCIPLE

RWS

Comparison of MT Engines at Foras na Gaeilge [EN-GA]

A 496-segment Test Set was created by Foras na Gaeilge, separate from all training data.

An automatic evaluation was conducted independently by DCU of four MT engines.



BLEU scores of four engines on a 496-segment test set provided by Foras na Gaeilge.

Evaluation of MT Output at Foras na Gaeilge [EN-GA]

Foras na Gaeilge

Two FnaG translators undertook Adequacy and Fluency evaluation of Iconic MT output on the 496 test segments, using a 4-point Likert scale. The questions were

- *How much of the information and meaning expressed in the source is conveyed accurately in the translation?*
- *How fluent is the translation?*

Measurement of inter-translator agreement:

Cohen's Kappa	Adequacy	Fluency
Non-weighted	0.009	0.011
Weighted	0.031	0.026

- *Generally low agreement between translators*
- *Translator 2 more strict – ratings 2-3, not 4*

Translators' Rating of Adequacy and Fluency

	Adequacy	Fluency
Average	3.57	3.36
Mode	4	4

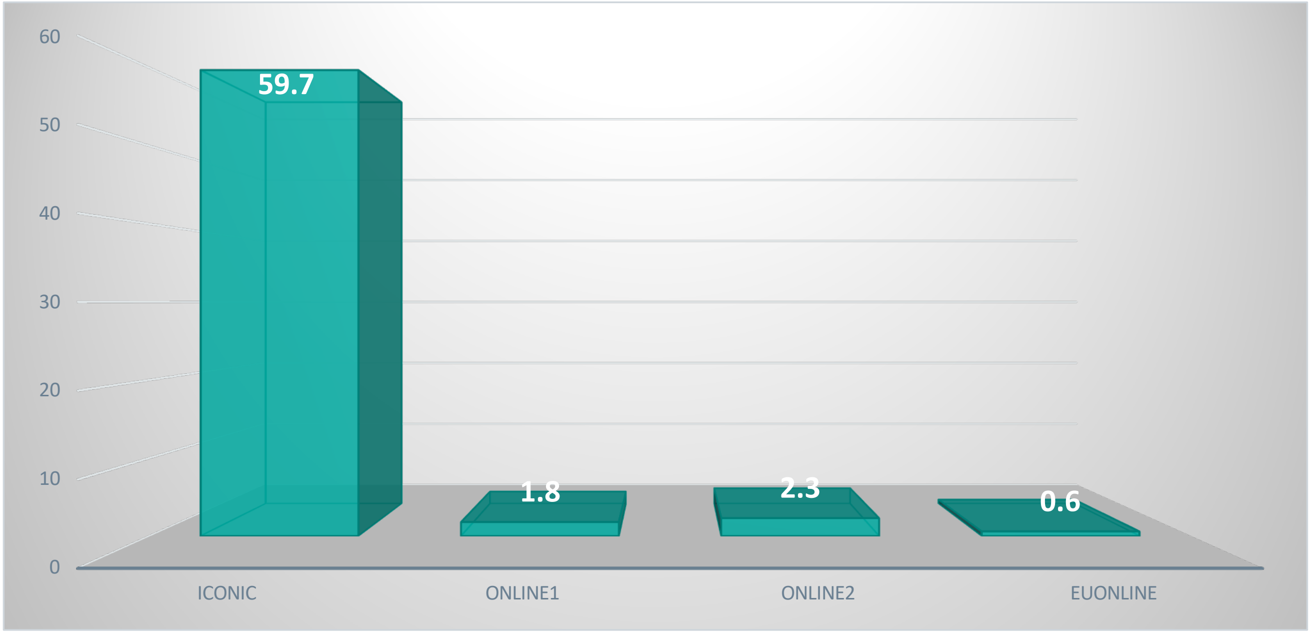
PRINCIPLE

RWS

Comparison of MT Engines at Met Office Iceland [IS-EN]

A 500-segment Test Set was created by Met Office Iceland, separate from all training data.

An automatic evaluation was conducted independently by DCU of four MT engines.



BLEU scores of four engines on a 500-segment test set provided by Met Office Iceland.

MT Post-Editing at Met Office Iceland [EN-IS]

Two Met Office translators undertook a Post-Editing exercise, each translator post-editing the entire 500 segment test set.

	Total Time	Avg. per Sentence
Translator 1	00:48:04	00:05.7
Translator 2	00:39:51	00:04.7

TER scores were calculated to compare similarity of MT output and PE result to the original reference translation, and HTER measured how much post-editing was performed on the MT output.

TER (Reference)	Translator1	Translator2
Iconic MT	22.7	22.7
PE	20.1	21.8

hTER (PE)	Translator1	Translator2
Iconic MT	12.9	5.9

- *Translator 2 performed fewer post-edits*

Deployment of MT to PRINCIPLE Early Adopter Users

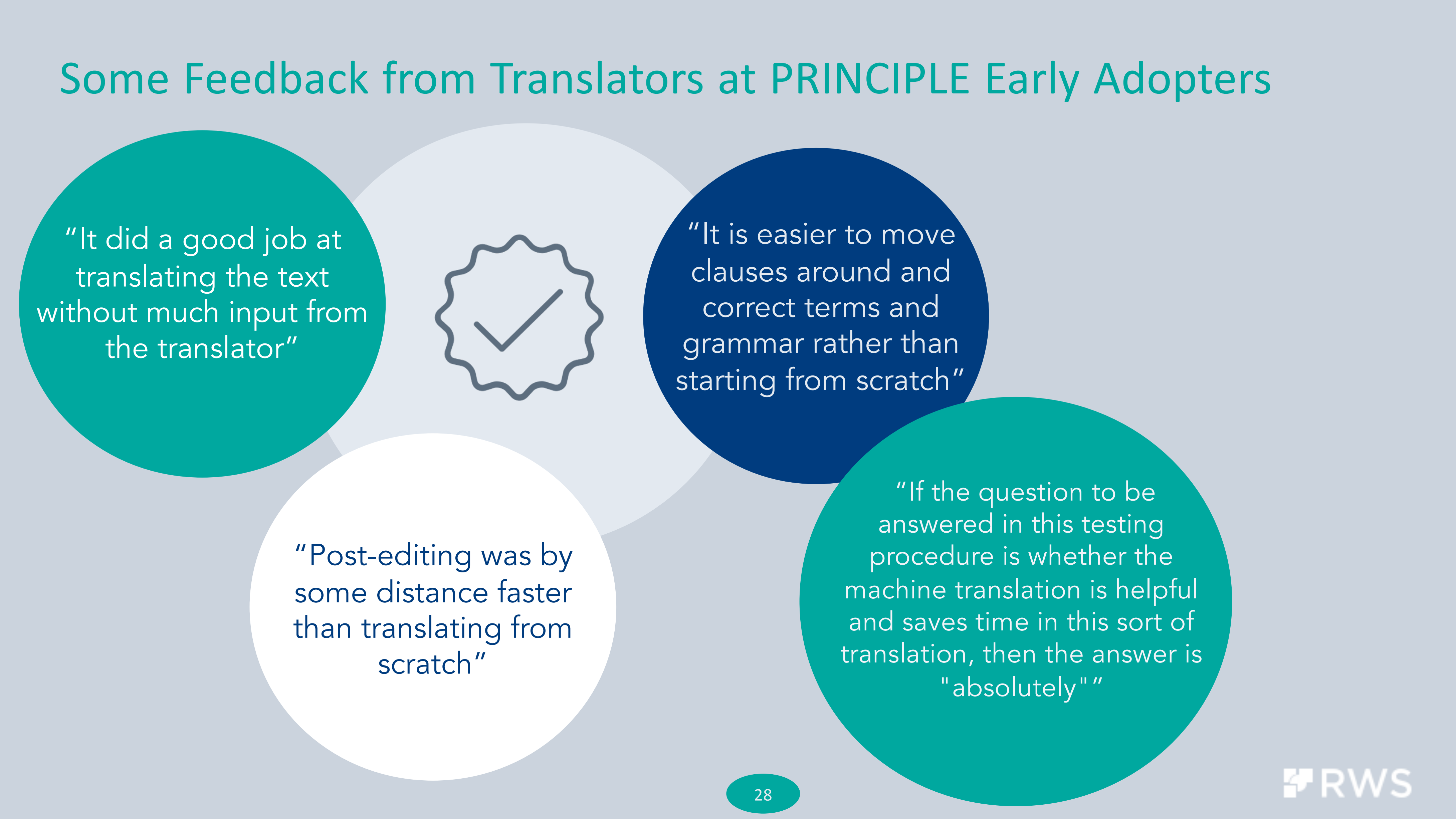
Each PRINCIPLE ’Early Adopter’ was set up with access to the MT model trained on their data for day-to-day use during the course of the project.

PRINCIPLE Early Adopters all work within the same use-case: MT to be used in conjunction with translator review / post-editing in the translation workflow.

Almost 1 million words have been processed through PRINCIPLE MT engines during the course of the project.

PRINCIPLE

RWS



Some Feedback from Translators at PRINCIPLE Early Adopters

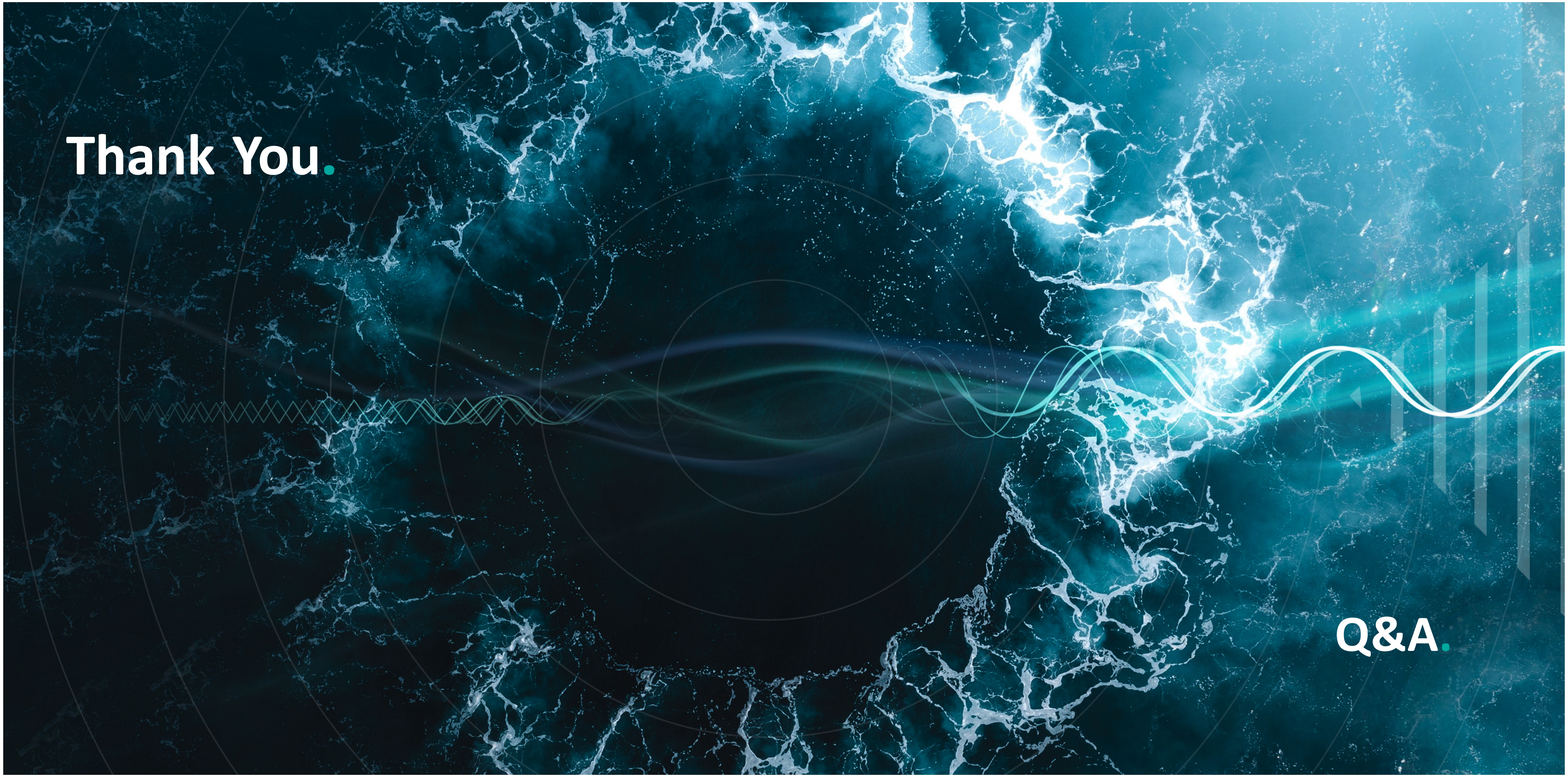
"It did a good job at translating the text without much input from the translator"

"It is easier to move clauses around and correct terms and grammar rather than starting from scratch"

"Post-editing was by some distance faster than translating from scratch"

"If the question to be answered in this testing procedure is whether the machine translation is helpful and saves time in this sort of translation, then the answer is "absolutely""

RWS



Thank You.

Q&A.

The work presented here is co-financed by the Connecting Europe Facility of the European Union

http://www.languageweaver.com
https://principleproject.eu

RWS