

A Different, Ethical MT is Possible:

English-Catalan Free/
Open-Source NMT

Vicent Briva-Iglesias
SFI Centre for Research Training in Digitally-Enhanced
Reality (D-REAL), Dublin City University

Overview

01

...

Understanding the Problem

Current context of MT.

02

...

Variables

MT engines and methodology.

03

...

Results

Relative ranking, quality,
and post-editing evaluation.

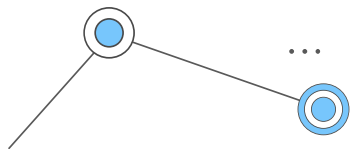
04

...

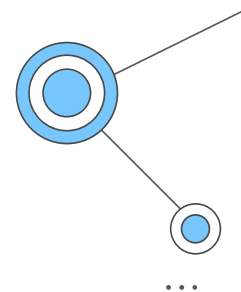
iMpacT

Effects and use-cases.





Understanding the Problem



01

Catalan context

Minoritized, stateless language. Low-resource.

02

NMT Requirements

Huge computational power (GPUs). Difficulty to find high-quality corpora for low-resource languages.

03

Literacy

You have the corpora. Now, how is an MT engine trained?

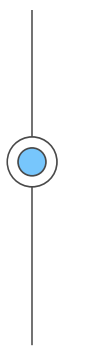
04

Data Privacy

Confidential information may be at stake.



What is the iMPacT of open-source MT for low-resource languages?

1. Which MT engine evaluated [Apertium, Softcatalà, Google] offers a higher translation quality?
 2. Which MT engine evaluated offers a bigger productivity increase when introducing it into a translation workflow?
 3. Can a free/open-source MT engine for a low-resource language beat the flagship MT engine for the English-Catalan language combination?
- 

Variables – MT Engines



Apertium

- Free/Open-Source RBMT engine
- Originally developed for close languages (e.g. ES-CA)



Softcatalà Translator

- Free/Open-Source EN<>CA NMT engine (OpenNMT)
- Trained with TMs from the Softcatalà project («in-domain»)



Google Translate

- Flagship of commercial MT
 - NMT from 2020
- Thousands of language combinations (including CA)

Variables – Text



[HomeAssistant.io](https://homeassistant.io)

- Open-source smart home software (GitHub)
- Preparation of the text with Okapi Framework
- Segments chosen randomly for the creation of the samples to be evaluated

Methodology – Human Evaluation 1

Relative Ranking

11 professional evaluators.

200 segments.

Rànquing de TA (Rank Comparison)

Source (English (United Kingdom))	
Start	
Current	This entity does not have a unique ID, therefore its settings cannot be managed from the UI.
Next	The {platform} integration is not loaded.
Target (Catalan)	
<input type="radio"/>	Aquesta entitat no té un ID únic, per tant la seva configuració no es pot gestionar des de la IU.
<input type="radio"/>	Aquesta entitat no té un ID únic, per tant, la seva configuració no es pot gestionar des de la interfície d'interès.
<input type="radio"/>	Aquesta entitat no té un únic ID, per tant no es poden abastar els seus paràmetres des del UI. (Info)
Comments	
<input type="text"/>	
Characters left: 500	

Methodology – Human Evaluation 2

Adequacy & Fluency

11 professional evaluators.

100 segments.

Precisió i fluïdesa S2, TA2

Source (English (United Kingdom))

Start

Current This service is run by our partner, a company founded by the founders of Home Assistant and Hass.io.

Next Go to the integrations page.

Target (Catalan)

Start

Current Aquest servei el gestiona el nostre soci, una empresa fundada pels fundadors de Home Assistant i Hass.io.

Next Vés a la pàgina d'integracions.

Fluency:

Incomprehensible Disfluent Good Flawless

[\(More Info\)](#)

Adequacy:

None Little Most Everything

[\(More Info\)](#)

Methodology – Human Evaluation 3

Post-Editing Evaluation

6 evaluators (2 groups of study: professionals & volunteers).

2 texts of 100 segments.

Information

Required Level of Quality: [Similar or equal to human translation](#)
Content Type: User Interface Text
Filename: PE_Sample1_TANS_taus_xlsx_empty_prod-qual.xlsx
Segment: 1 of 100

Source: English (United Kingdom)

Start

Current This entity does not have a unique ID, therefore its settings cannot be managed from the UI.

Next The {platform} integration is not loaded.

Target: Catalan

Start

Current Aquesta entitat no té un ID únic, per tant la seva configuració no es pot gestionar des de la IU.

PAUSE

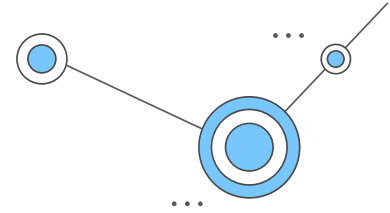
NEXT

Or Press Enter

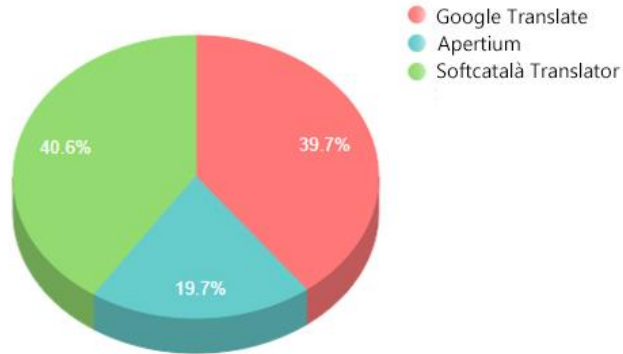
Post-Editing Evaluation (Explanation)

	Text 1, Engine 1	Text 1, Engine 2	Text 2, Engine 1	Text 2, Engine 2
Evaluator 1	✓	✗	✗	✓
Evaluator 2	✗	✓	✓	✗
Evaluator 3	✓	✗	✗	✓
Evaluator 4	✗	✓	✓	✗

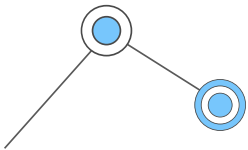
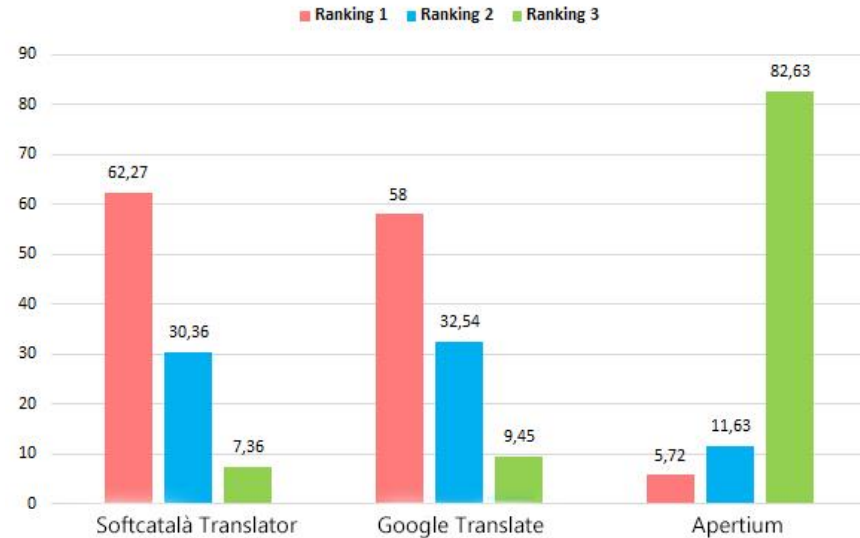
Results – MT Ranking



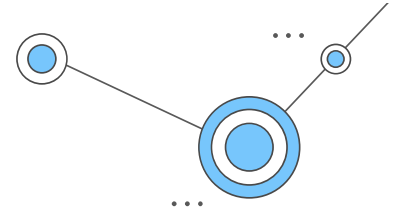
% of times an engine has received Ranking 1 evaluation



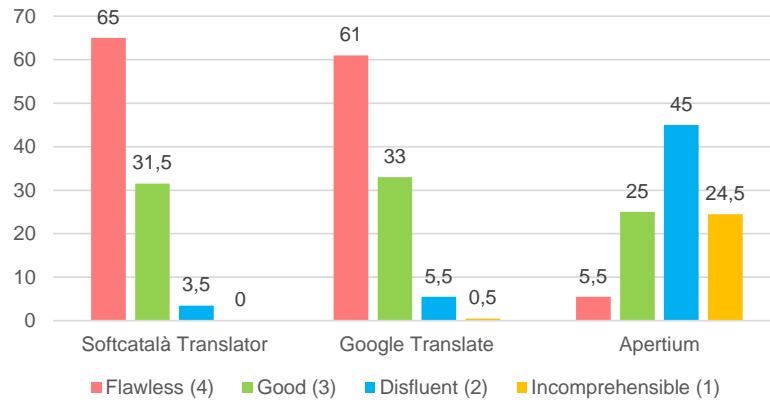
Ranking distribution per engine (in %)



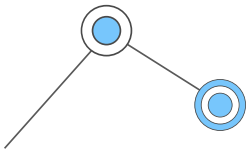
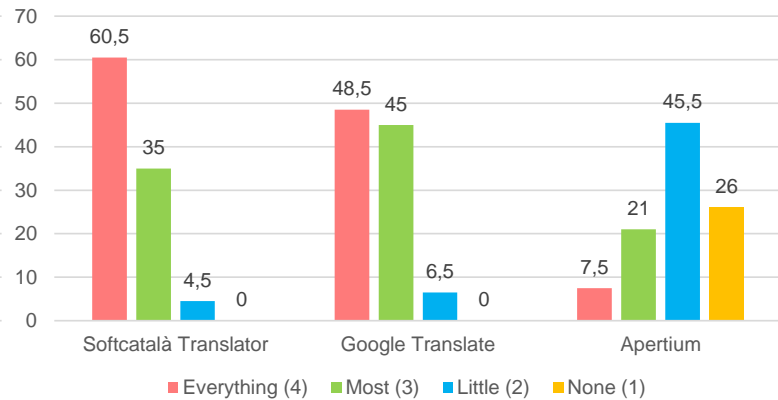
Results – Fluency & Adequacy



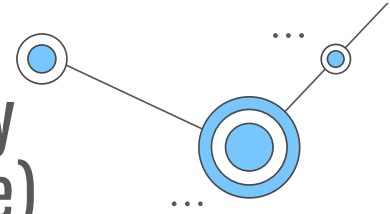
Fluency (in %)



Adequacy (in %)



Results – Post-Editing Productivity (group of study 1: Softcatalà–Google)



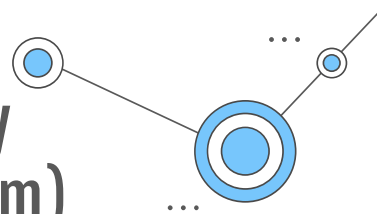
	Softcatalà Translator	Google Translate
PE Time (s)	Median 3909.07	Median 4131.64
Edit Distance* (segment)	9.79	10.35

222.563 seconds of difference; 5.69% productivity increase

		1-5 words	6-15 words	16 or >16 words
PE Time (s)	Softcatalà	Median 8.15	Median 18.44	Median 34.08
	Google	9.41	20.08	33.67

Edit distance* (seg.)	Softcatalà	5.34	11.53	9.79
	Google	12.22	9.31	11.20

Results – Post-Editing Productivity (group of study 2: Softcatalà-Apertium)



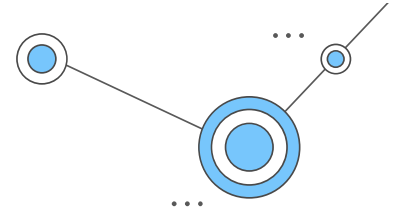
	Softcatalà Translator	Apertium
PE Time* (s)	Median 1859.51	Median 3743.41
Edit Distance* (segment)	6.81	24.85

1883.89 seconds of difference; 101.31 % productivity increase

		1-5 words	6-15 words	16 or >16 words
PE Time* (s)	Softcatalà	Median 5.95	Median 14.18	Median 25.83
	Apertium	14.11	28.64	55.70

Edit distance* (seg.)	Softcatalà	6.21	10.73	10.11
	Apertium	40.65	37.76	36.15

iMPact and Effects

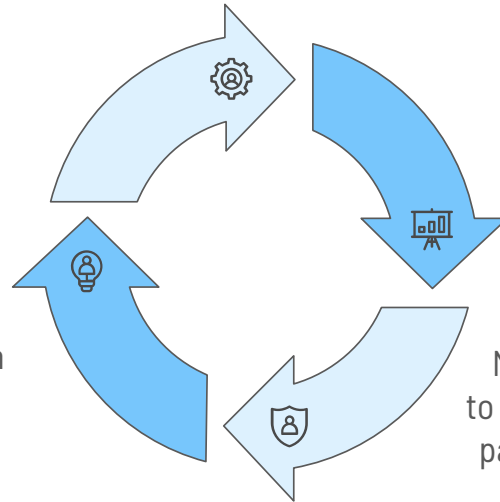


Normalisation

Low-resource languages gain presence on the Internet, society, etc.

Data Privacy

Confidential information is preserved.

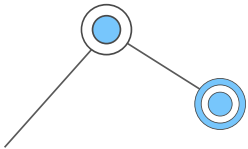


Language Diversity

Avoid language shifts to predominant languages. And fosters language literacy.

Crisis Scenarios

Multilingual communication to reach everyone, e.g. COVID pandemic, natural disasters.



Thanks!

Do you have any questions?

Vicent Briva-Iglesias
D-REAL, Dublin City University
vicent.brivaiglesias2@mail.dcu.ie
@VicentBriva

CREDITS: This presentation template was created by [Slidesgo](#), including icons by [Flaticon](#), infographics & images by [Freepik](#) and illustrations by [Stories](#)

