tarjama WORDSMITHS

Real-World Custom NMT for Arabic

A Data-Centric Approach

August 2021

Dr. Rebecca Jonsson – Head of AI Products
Ruba Jaikat – Applied ML Scientist Lead

Information Security Management System ISO 27001 Certified

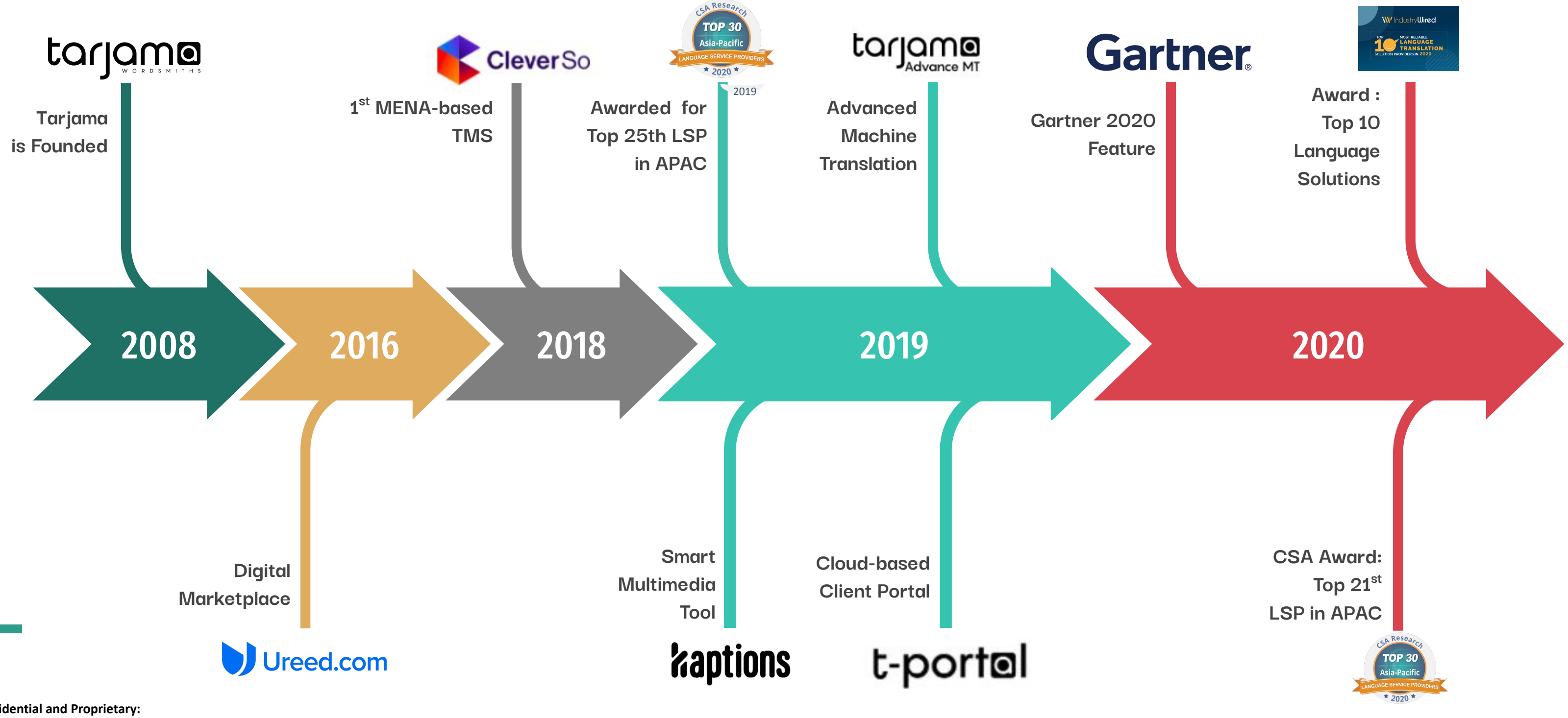
CSA Research TOP 30 Asia-Pacific LANGUAGE SERVICE PROVIDERS 2020

ERNST & YOUNG ENTREPRENEUR OF THE YEAR 2018

SME TECH COMPANY OF THE YEAR 2019

12 Years of Innovation in Language Technology

A Language Service Provider in MENA region turning into a Language Tech Provider



2008

Tarjama is Founded

2016

Digital Marketplace

Ureed.com

2018

1st MENA-based TMS

CleverSo

2019

Awarded for Top 25th LSP in APAC

Advanced Machine Translation

tarjama Advance MT

Smart Multimedia Tool

kaptions

Cloud-based Client Portal

t-portal

2020

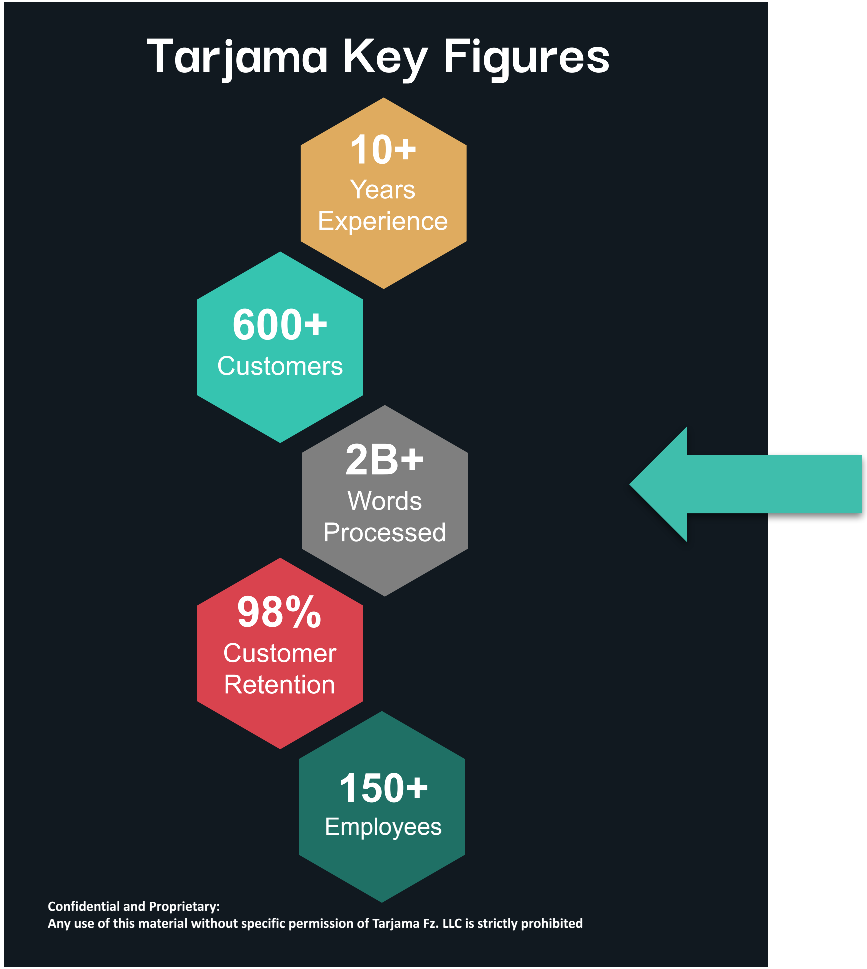
Gartner 2020 Feature

Gartner

Award : Top 10 Language Solutions

CSA Award: Top 21st LSP in APAC

Confidential and Proprietary:
Any use of this material without specific permission of Tarjama Fz. LLC is strictly prohibited



Tarjama Key Figures

10+ Years Experience

600+ Customers

2B+ Words Processed

98% Customer Retention

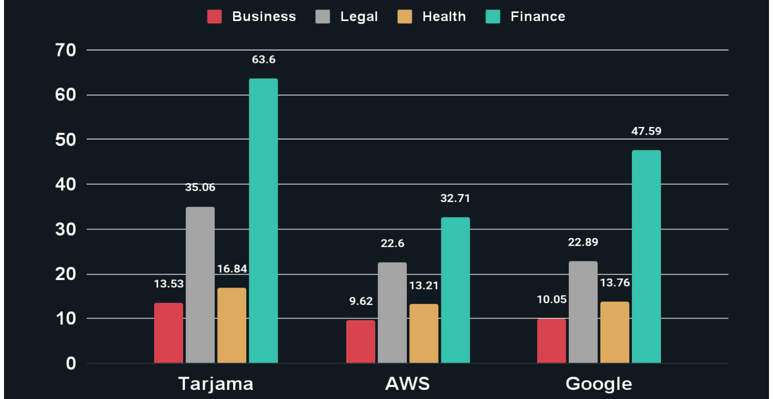
150+ Employees

1. Female-led LSP transforming into a language technology company.
2. Localization, Translation, Interpreting, Subtitling and Content creation services.
3. Dominant in MENA region focusing on Arabic language and dialects.
4. Proprietary TMS system with focus on Arabic support.
5. Proprietary NMT system for EN-AR

Comparison Evaluation

(EN→ AR)

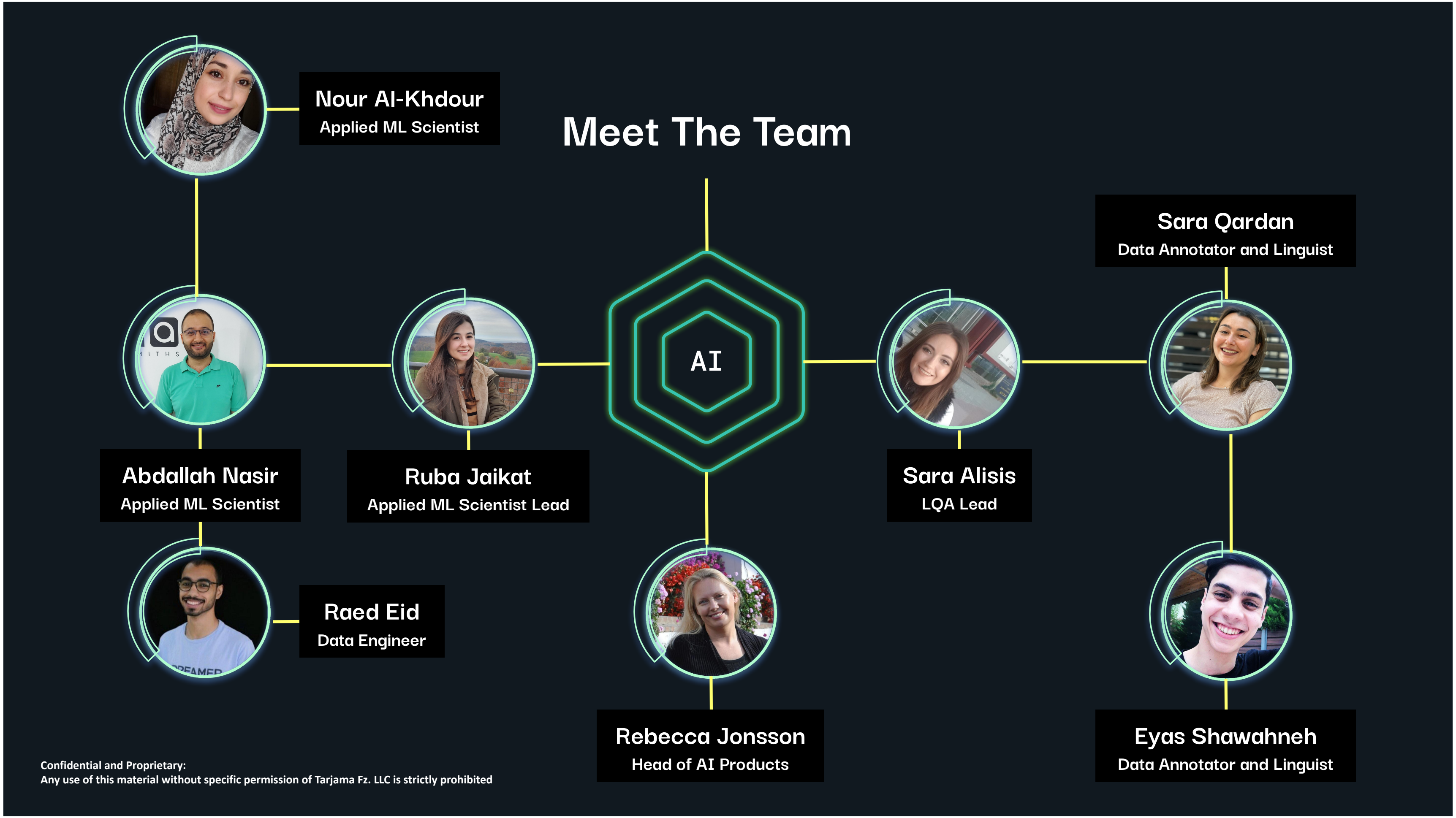
- Comparison Evaluation Set comprises of 120 segments with a total of 2.6k word count.
- Domain segments count is distributed as follows; Business: 38, Legal: 20, Health: 30, Finance: 40.
- Text is collected from online articles.



BLEU Scores on the Comparison Evaluation Set using Tarjama, AWS, and Google MT Engines.

Tarjama NMT Engine

- Tarjama NMT Engine development started late 2019.
- Tarjama NMT Engine is trained on high-quality Data translated by expert linguists.
- Tarjama Data covers various business domains, including: Legal, Consultancy, Health, Finance, Marketing, E-Commerce, Medical, Culture, News, Politics, Technology, Entertainment and more.
- Gold nuggets of external publicly available datasets are extracted and used to further enrich the engine.
- Currently, the use of Tarjama NMT within the Translation process reaches up to 35%.
- Productivity tests show that post-editing Tarjama NMT output saves at least 40% of the translator time.



Meet The Team

AI

Nour Al-Khdour
Applied ML Scientist

Abdallah Nasir
Applied ML Scientist

Raed Eid
Data Engineer

Ruba Jaikat
Applied ML Scientist Lead

Rebecca Jonsson
Head of AI Products

Sara Alisis
LQA Lead

Sara Qardan
Data Annotator and Linguist

Eyas Shawahneh
Data Annotator and Linguist

Confidential and Proprietary:
Any use of this material without specific permission of Tarjama Fz. LLC is strictly prohibited

Tailored NMT models

- Going beyond Custom MT by tailoring a NMT model fit for the needs of a customer.
- Data-centric approach selecting the gold nuggets of their data and considering translation guidelines.
- Model that performs best-in-class on the customer data.
- Generalizes well on other data sets.
- Human Evaluation of candidate models to select a high-quality model.

Tailored NMT Development Cycle

01

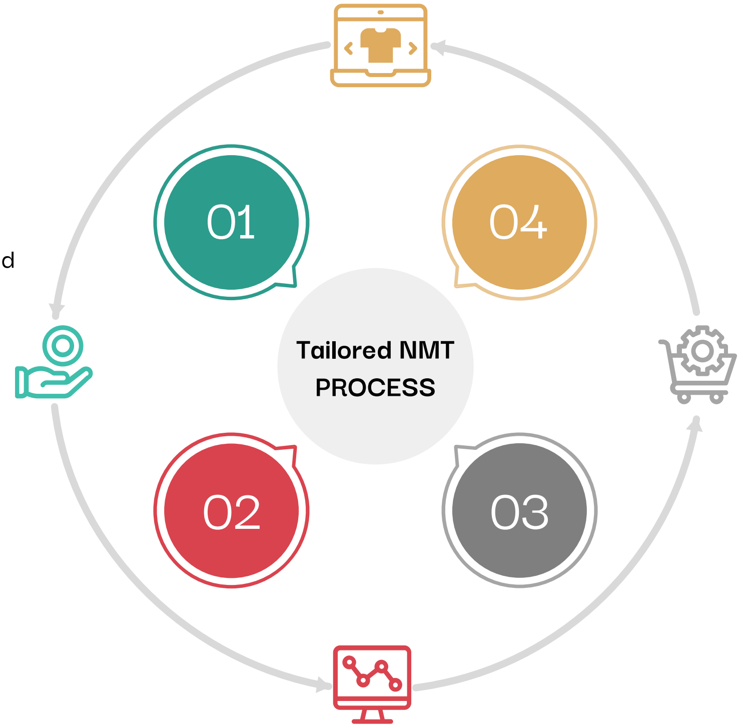
Data Acquisition & Analysis

Client Data received then analyzed by Linguistic QA Experts

02

Data Preprocessing & Filtering

Client Data run through Tarjama Data pipeline for preprocessing and filtering (selecting gold nuggets)



01

04

Tailored NMT PROCESS

02

03

04

Model Adaptation

Experimenting, fine-tuning, analyzing, and evaluating the MT engine and its performance with client data

03

Add External Data

Carefully selecting out-of-domain data to add together with client data with the purpose of building a robust tailored MT engine that generalizes to other data

E-commerce data

”Stylized collectable stands 3 3/4 inches tall, perfect for any Harry Potter fan”

”The textured fabric truly brings this T-Rex to life“

”rectangular sunglasses ar 8069 5447/11“

”256GB NVMe SSD + 1TB (7200Rpm)“

”waterproof sun protection full car cover for gmc k15/k1500 pickup 1971-67“

”This Speed Cube Bundle (2x2x2 cube, 3x3x3 cube, pyramid 3x3x3 cube) is the classic color-matching puzzle, perfect for reducing stress & exercising your brain & improving memory & practicing hands-on dexterity skills”

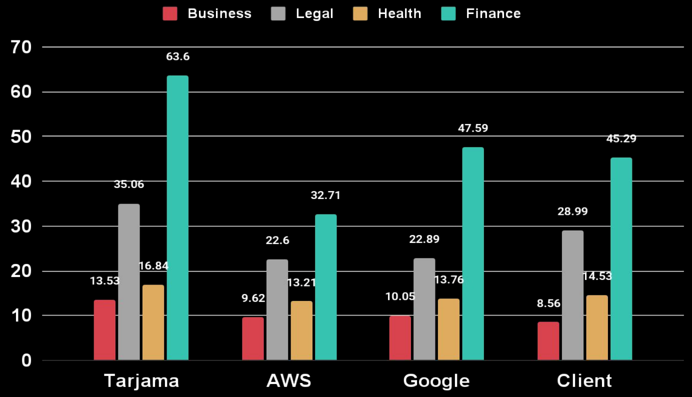
”With the 144 Hz full HD, 1920 x 1080 display, on-screen action is incredibly smooth and fluid”

”2 in 1 ipad air case cover smart case cover with magnetic auto wake & sleep feature trifold stand for apple ipad air (ipad 5) tablet“

Confidential and Proprietary:
Any use of this material without specific permission of Tarjama Fz. LLC is strictly prohibited

Tailoring NMT for an e-commerce client

Dataset: 3M bilingual (EN→ AR) segments - high quality

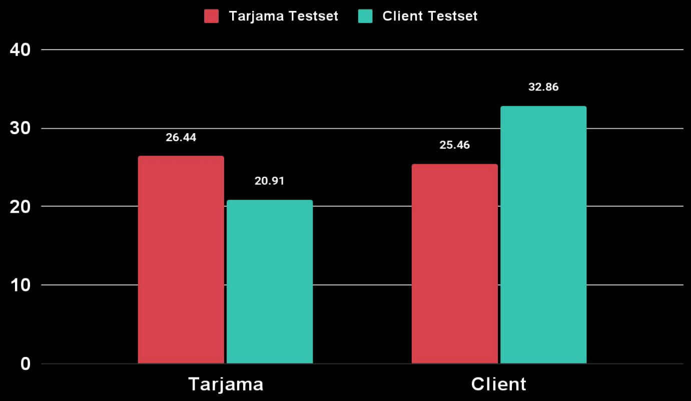


BLEU Scores on the Comparison Evaluation Set using Tarjama, AWS, Google, and Tailored MT Engines.

tarjama
WORDSMITHS

Tailoring NMT for an e-commerce client

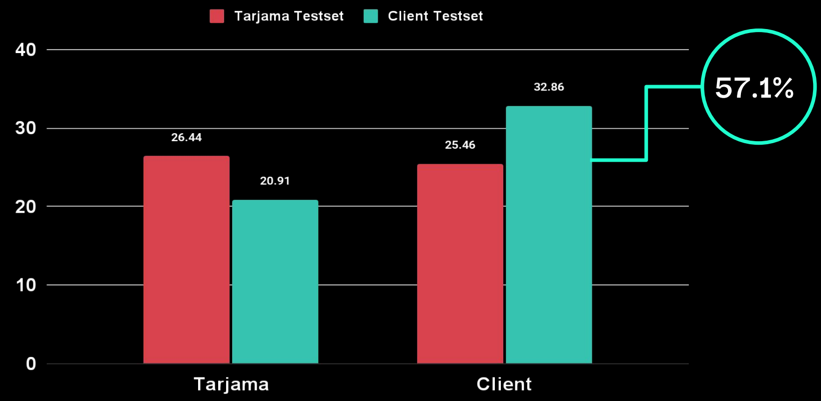
Dataset: 3M bilingual (EN→ AR) segments - high quality



BLEU Scores on the Tarjama (5k) and Client (5k) Testing sets using Tarjama Generic and Client's Tailored MT engines.

Tailoring NMT for an e-commerce client

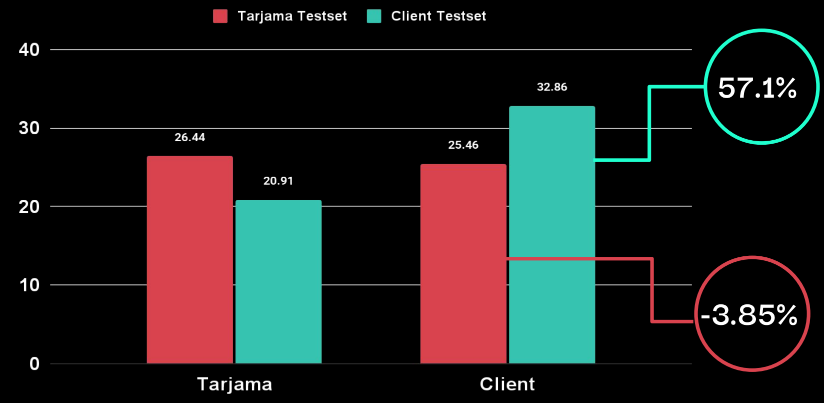
Dataset: 3M bilingual (EN→ AR) segments - high quality



BLEU Scores on the Tarjama (5k) and Client (5k) Testing sets using Tarjama Generic and Client's Tailored MT engines.

Tailoring NMT for an e-commerce client

Dataset: 3M bilingual (EN⟶ AR) segments - high quality

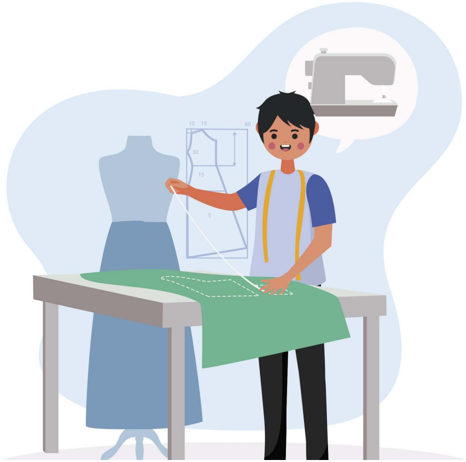


BLEU Scores on the Tarjama (5k) and Client (5k) Testing sets using Tarjama Generic and Client's Tailored MT engines.

Manual Evaluation

Adapted MQM approach

- Manual Evaluation of 500 segments (4212 words) translated with the tailored NMT
  - 86% of the translations considered OK, Good or Perfect. Minor review.
  - Most common error: 4.5 % mistranslations



MT Quality	Distribution
Perfect MT translation	63%
Good MT translation (minor errors)	1.6%
OK translation (a few errors)	21.8%
Bad translation	11.8%
Nonsense translation	1.8%

Source: English

The luxurious-feeling moisturizer immediately leaves skin hydrated and softens the look of fine lines and wrinkles

Brow line frame sunglasses
257-17c

lcd backlight display for clear and fast reading of measurement data

Materialsilicone

MT Target: Arabic

مرطب ذو ملمس فاخر يترك البشرة رطبة
على الفور وينعم مظهر الخطوط الدقيقة
والتجاعيد

نظارة شمسية بإطار يغطي الحاجب طراز
17C-257

شاشة LCD بإضاءة خلفية لقراءة بيانات القياس
بشكل واضح وسريع

مصنوع من السيليكون

Confidential and Proprietary:
Any use of this material without specific permission of Tarjama Fz. LLC is strictly prohibited

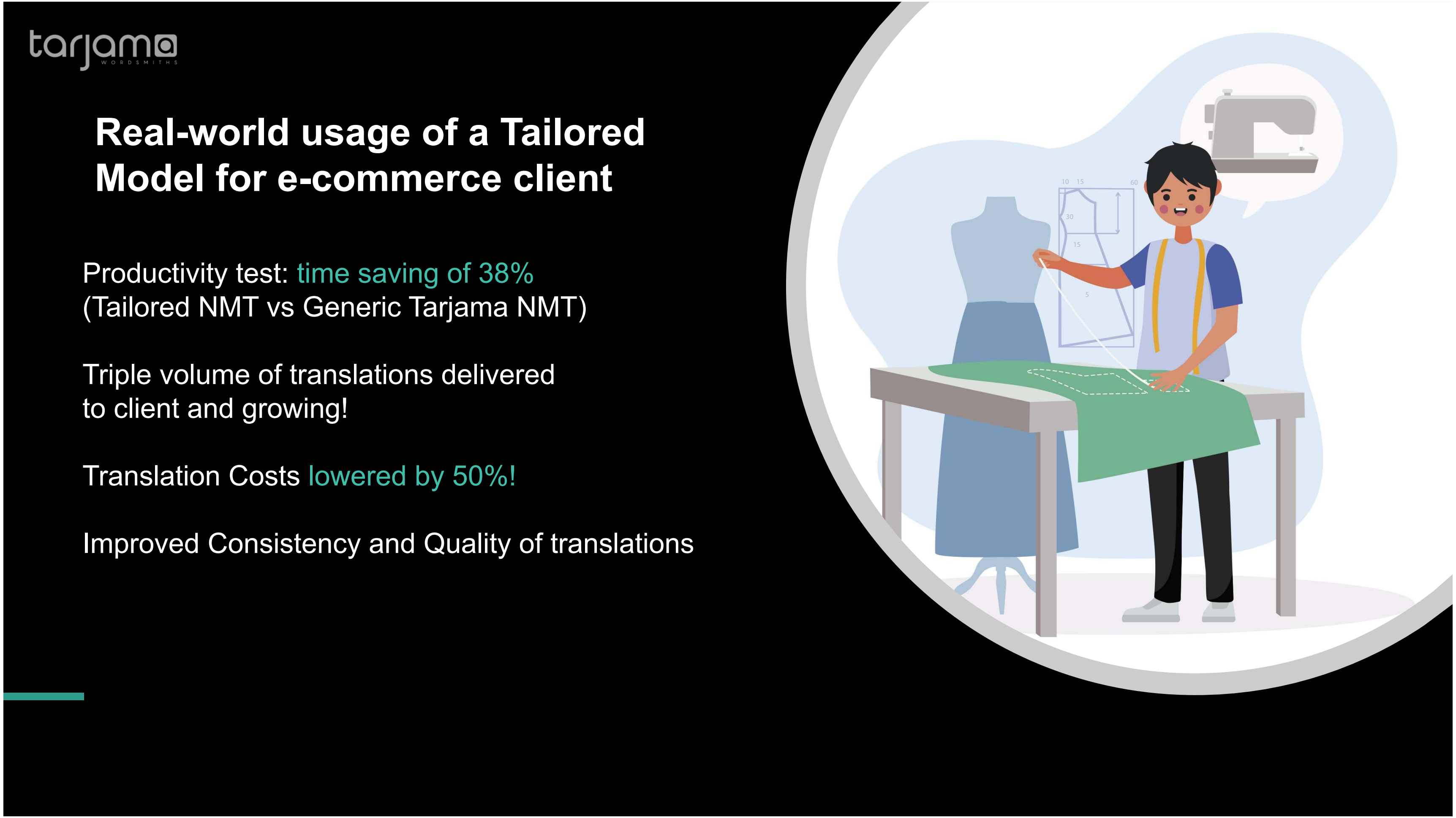
tarjama WORDSMITHS

Real-world usage of a Tailored Model for e-commerce client

Translation of e-commerce data from English to Arabic using Tarjama’s TMS system for an e-commerce client.

60-90 Translators (post-editors) in-house and freelancers.

Tailored NMT model used for pre-translation and translators performing post-editing and transcreation.



tarjama
WORDSMITHS

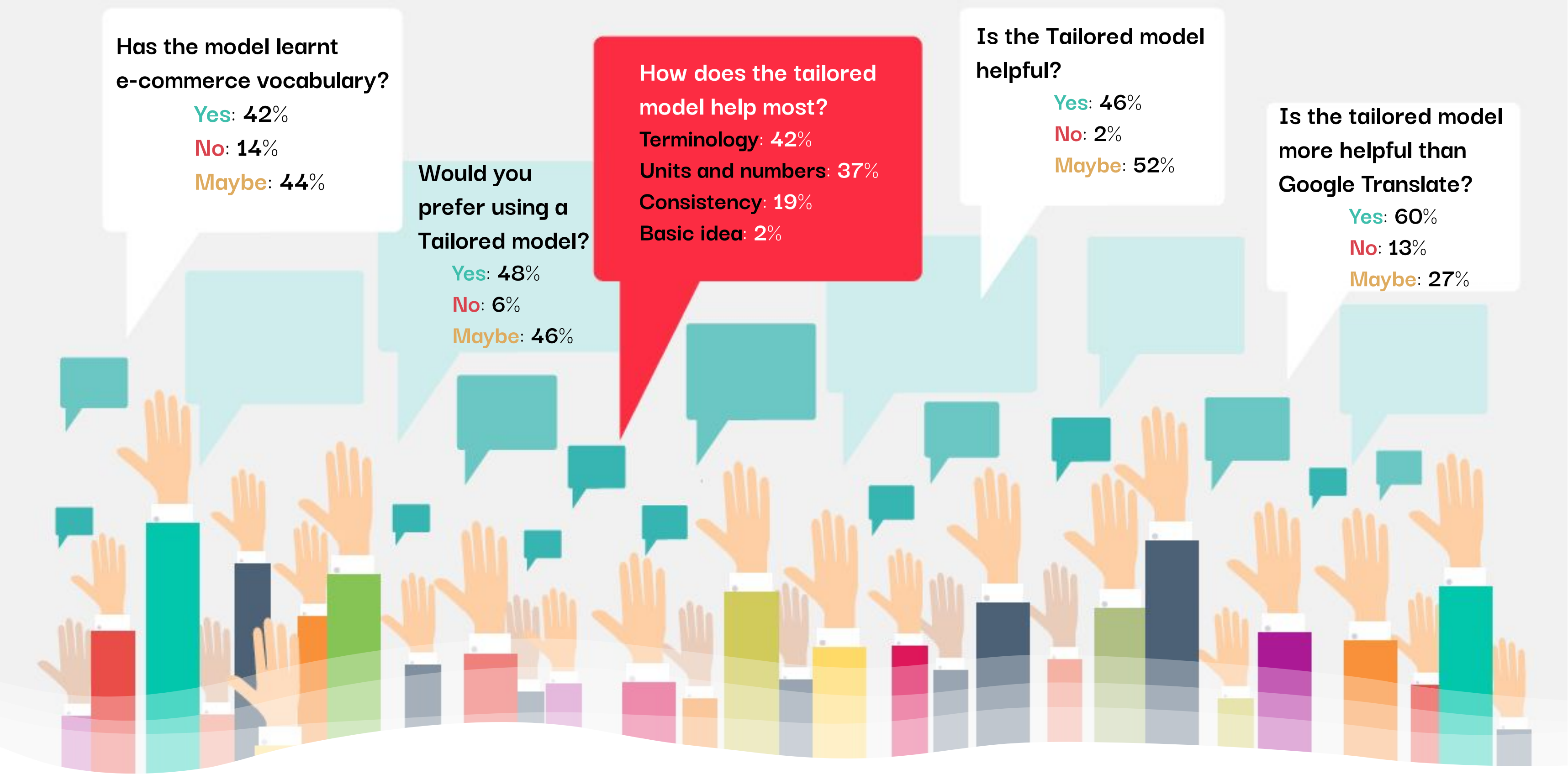
Real-world usage of a Tailored Model for e-commerce client

Productivity test: time saving of 38%
(Tailored NMT vs Generic Tarjama NMT)

Triple volume of translations delivered to client and growing!

Translation Costs lowered by 50%!

Improved Consistency and Quality of translations



Has the model learnt e-commerce vocabulary?

Yes: 42%
No: 14%
Maybe: 44%

Would you prefer using a Tailored model?

Yes: 48%
No: 6%
Maybe: 46%

How does the tailored model help most?

Terminology: 42%
Units and numbers: 37%
Consistency: 19%
Basic idea: 2%

Is the Tailored model helpful?

Yes: 46%
No: 2%
Maybe: 52%

Is the tailored model more helpful than Google Translate?

Yes: 60%
No: 13%
Maybe: 27%

What did the translators think?

□ Survey with 50 translators
□ 65% has experience in post-editing

tarjama
WORDSMITHS

thank you
شكراً جزيلاً
Merci beaucoup
ありがとう
धन्यवाद
خیلی ممنونم

www.tarjama.com