Evaluating Gender Bias in Machine Translation

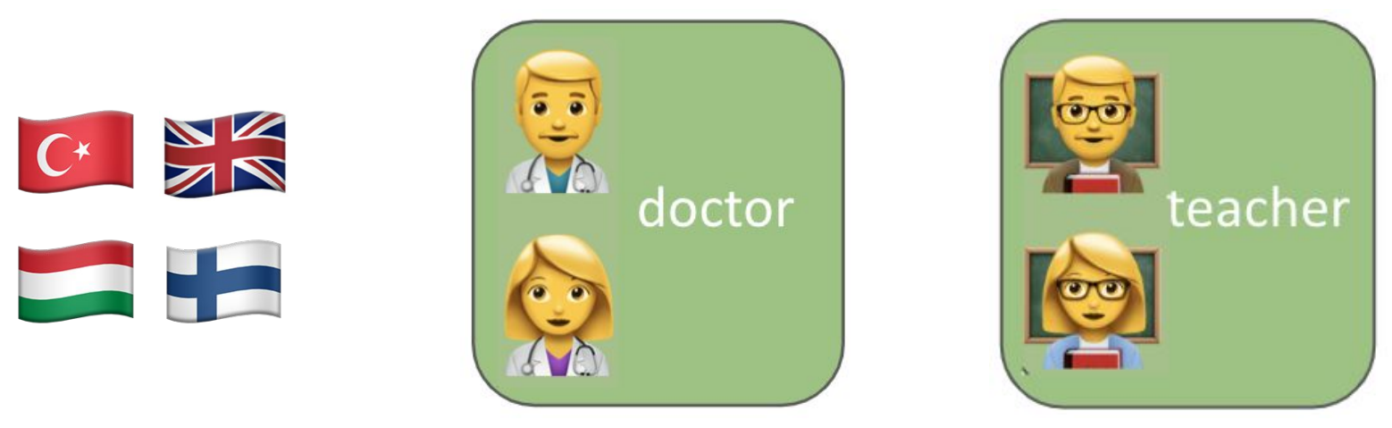
Gabriel Stanovsky, Noah Smith and Luke Zettlemoyer

Grammatical Gender

Some languages encode grammatical gender

doctor
maestro
doctora
maestra

… others don't



Translating Gender

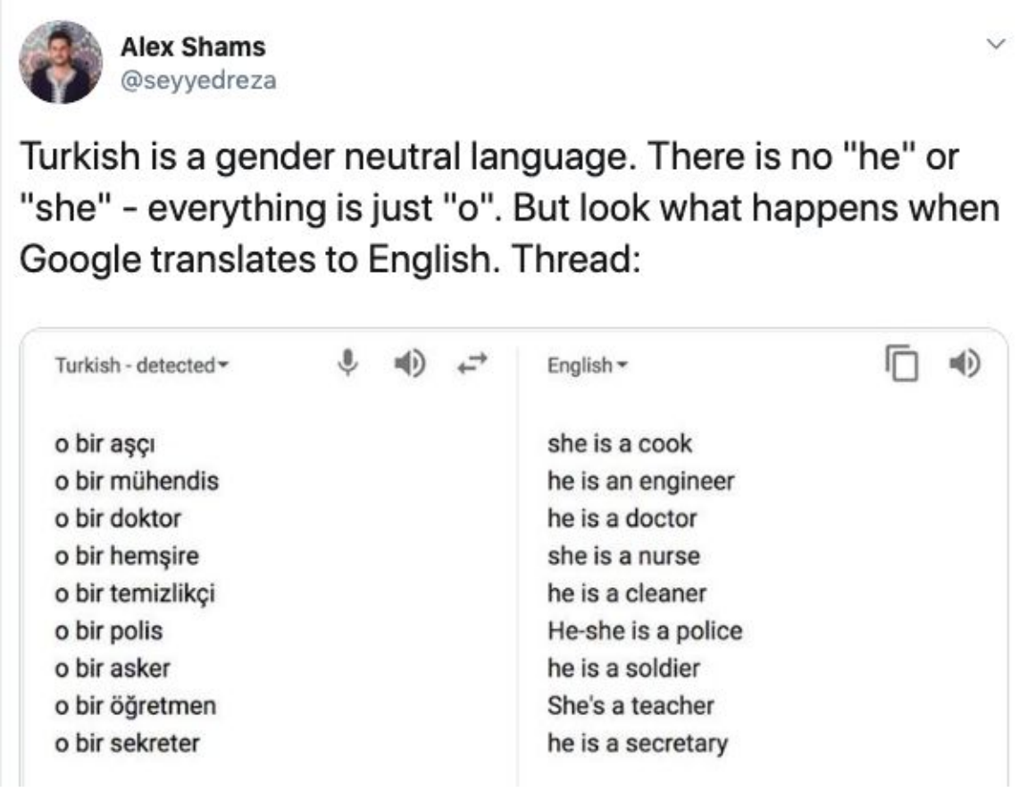
Variations in gender mechanisms across languages **prohibit one-to-one translations**

The doctor asked the nurse to help her in the procedure.

La doctora le pidió a la enfermera que le ayudara con el procedimiento.

→ Translating between English Spanish, for example, requires **making decisions about gender**

Is MT Gender-Biased?

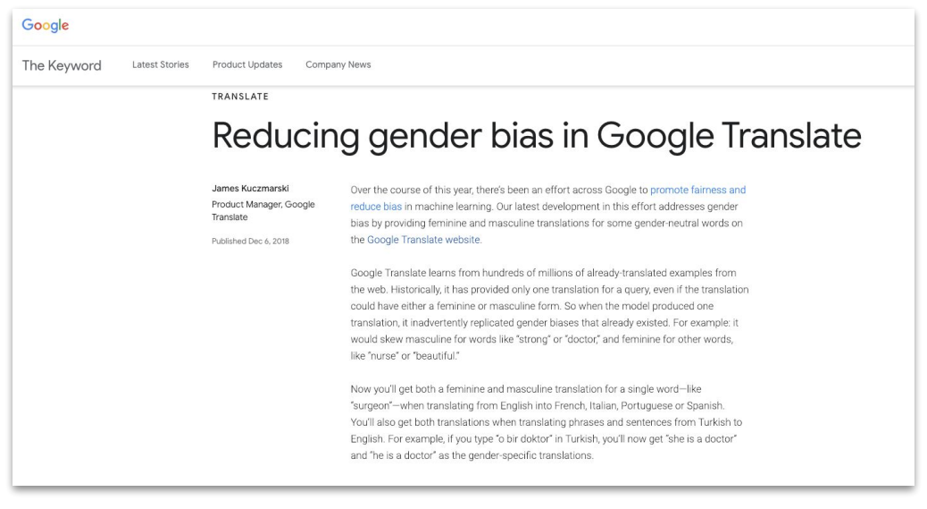


Alex Shams
@seyyedreza

Turkish is a gender neutral language. There is no "he" or "she" - everything is just "o". But look what happens when Google translates to English. Thread:

o bir aşçı	she is a cook
o bir mühendis	he is an engineer
o bir doktor	he is a doctor
o bir hemşire	she is a nurse
o bir temizlikçi	he is a cleaner
o bir polis	He-she is a police
o bir asker	he is a soldier
o bir öğretmen	She's a teacher
o bir sekreter	he is a secretary

- **Anecdotal evidence** that popular MT services resort to stereotypical role assignments



Reducing gender bias in Google Translate

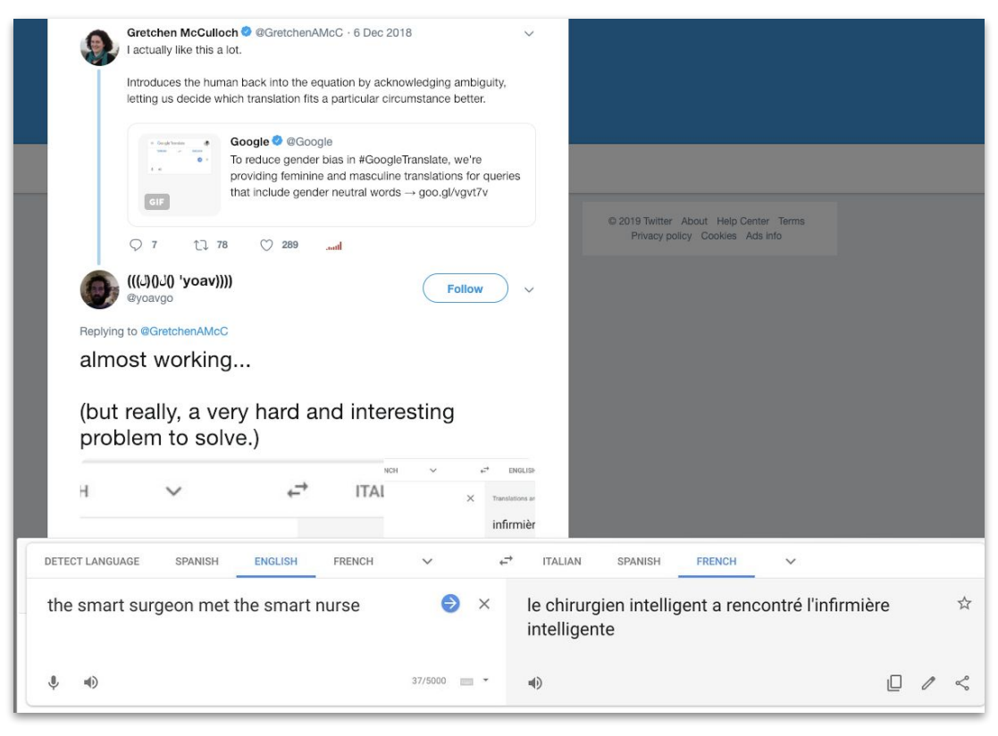
James Kuczmarski
Product Manager, Google Translate
Published Dec 6, 2018

Over the course of this year, there's been an effort across Google to promote fairness and reduce bias in machine learning. Our latest development in this effort addresses gender bias by providing feminine and masculine translations for some gender-neutral words on the Google Translate website.

Google Translate learns from hundreds of millions of already-translated examples from the web. Historically, it has provided only one translation for a query, even if the translation could have either a feminine or masculine form. So when the model produced one translation, it inadvertently replicated gender biases that already existed. For example: it would skew masculine for words like "strong" or "doctor," and feminine for other words, like "nurse" or "beautiful."

Now you'll get both a feminine and masculine translation for a single word—like "surgeon"—when translating from English into French, Italian, Portuguese or Spanish. You'll also get both translations when translating phrases and sentences from Turkish to English. For example, if you type "o bir doktor" in Turkish, you'll now get "she is a doctor" and "he is a doctor" as the gender-specific translations.

- Google Translate introduces a feature to control gender inflection for individual words



Gretchen McCulloch @GretchenAMcC · 6 Dec 2018
I actually like this a lot.

Introduces the human back into the equation by acknowledging ambiguity, letting us decide which translation fits a particular circumstance better.

Google @Google
To reduce gender bias in #GoogleTranslate, we're providing feminine and masculine translations for queries that include gender neutral words → goo.gl/vgvt7v

7 78 289

(((ل()ل() 'yoav)))) @yoavgo
Replying to @GretchenAMcC

almost working...

(but really, a very hard and interesting problem to solve.)

the smart surgeon met the smart nurse

le chirurgien intelligent a rencontré l'infirmière intelligente

- Not yet a viable solution for complete sentences

Estimating MT Gender Accuracy

In: MT model + target language
Out: Accuracy score for gender translation

Quality estimated at > 85% vs. 90% IAA
Doesn't require reference translations!

English source texts

- Winogender (Rudinger et al., 2018), WinoBias (Zhao et al., 2018)
  - 3888 English sentences designed to test gender bias in coreference resolution
  - Following the Winograd schema

The doctor asked the nurse to help her in the procedure.

The doctor asked the nurse to help him in the procedure.

→ **Observation:** Also useful for evaluating MT gender-bias!
  - Equally split between *stereotypical* and *non-stereotypical* role assignments
  - Gold annotations for gender

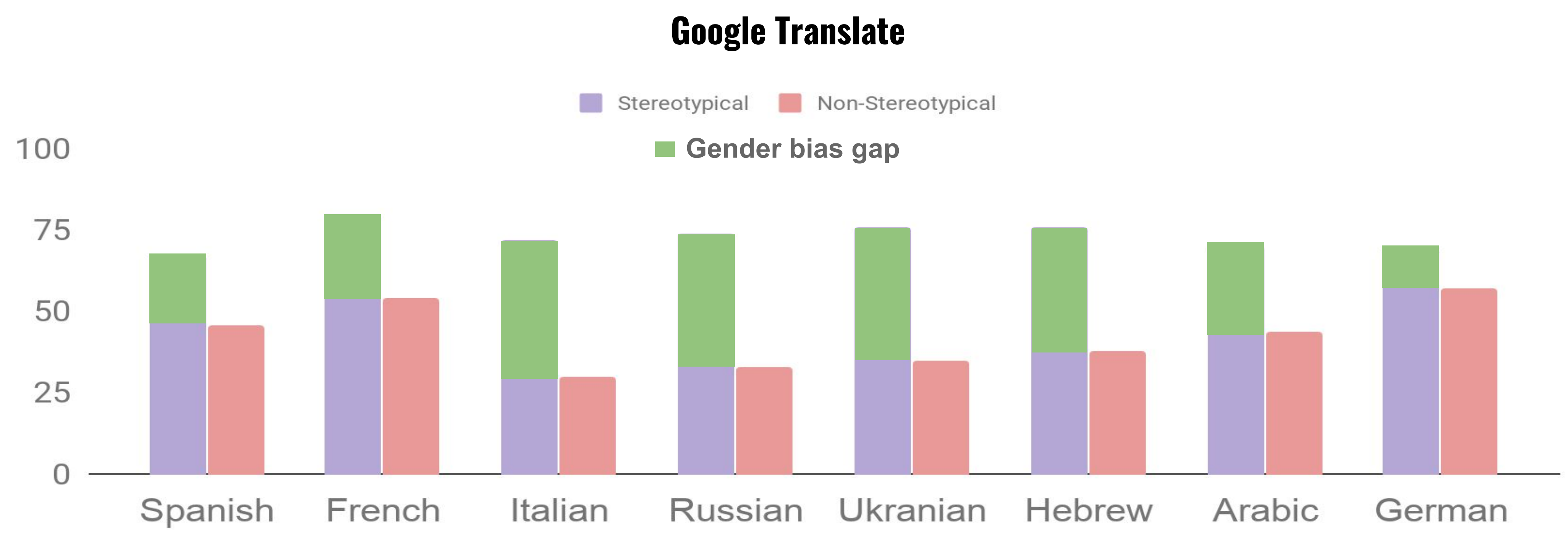
Automatic evaluation

1. **Translate the coreference bias datasets**
   To target languages with grammatical gender
2. **Align between source and target**
   Using fast align (Dyer et al., 2013)
3. **Identify gender in target language.**
   Using off-the-shelf morphological analyzers or simple heuristics in the target languages

The doctor asked the nurse to help her in the procedure.

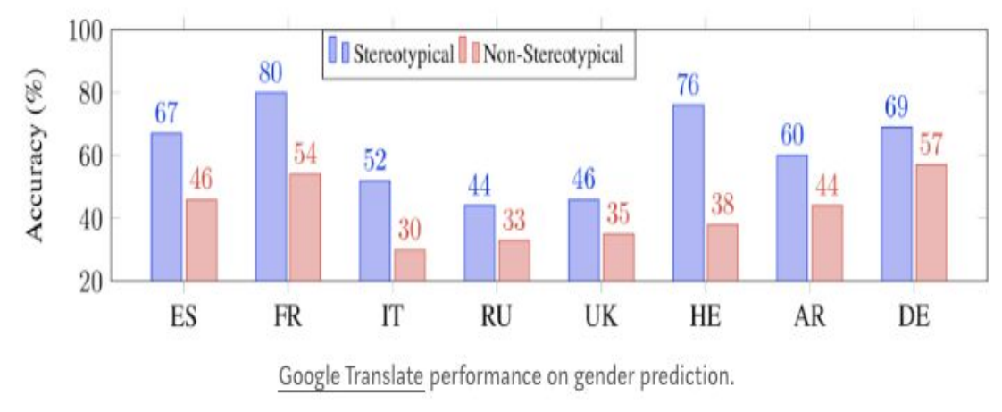
La doctora le pidió a la enfermera que le ayudara con el procedimiento.

Main Findings

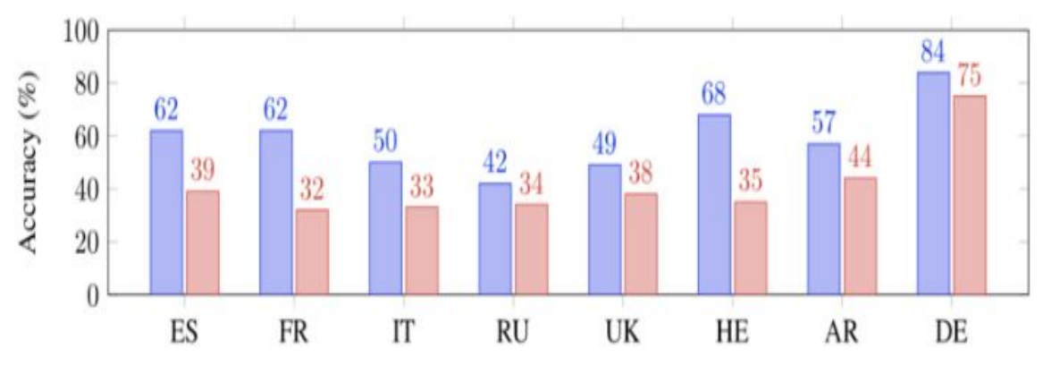


All models were significantly gender-biased

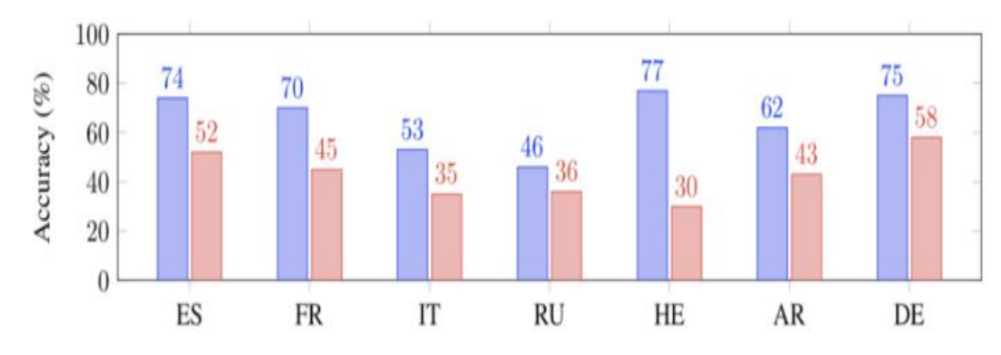
- MT struggles with non-stereotypical roles across languages and systems, **often doing significantly worse than random coin-flip**
- Academic models (Ott et al., 2018; Edunov et al., 2018) exhibit similar behavior



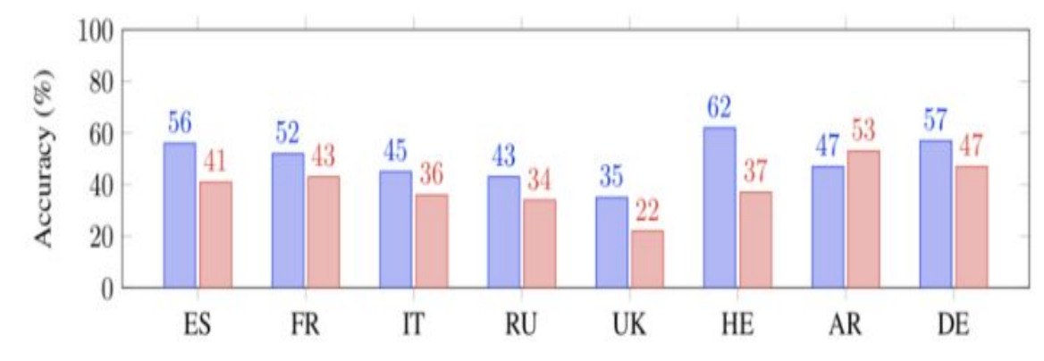
Google Translate performance on gender prediction.



Microsoft Translator



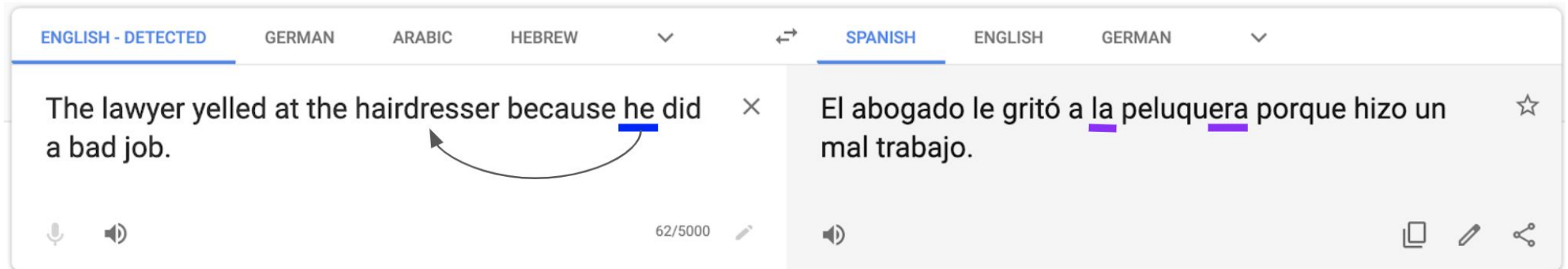
Amazon Translate



SYSTRAN

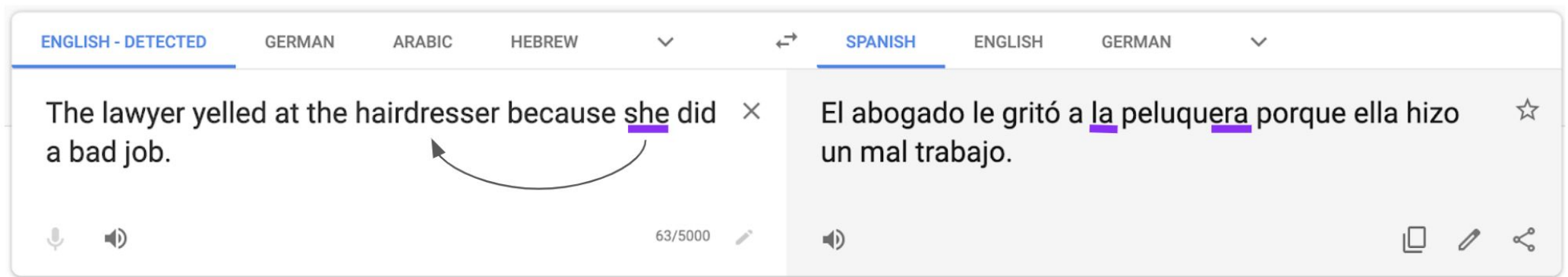
Gendered adjectives affect translation

- Black-box injection of gendered adjectives:
  - *the **pretty** doctor asked the nurse to help her in the operation*
  - *the **handsome** nurse asked the doctor to help him in the operation*
- Improved performance for most tested languages and models
  - + 10% on Spanish and Russian [mean +8.6%]
- **Requires oracle coreference resolution!**
  - Attests to the relation between coreference resolution and machine translation
  - Improvement in coreference resolution directly improves the accuracy of gender translation



The lawyer yelled at the hairdresser because he did a bad job.

El abogado le gritó a la peluquera porque hizo un mal trabajo.



The lawyer yelled at the hairdresser because she did a bad job.

El abogado le gritó a la peluquera porque ella hizo un mal trabajo.