Translated Wikipedia Biographies

English -> Spanish (516 KB)
English-> German (517 KB)

The Translated Wikipedia Biographies dataset has been designed to evaluate gender accuracy in long text translations (multiple sentences or passages). The set has been designed to analyze common gender errors in machine translation like incorrect gender choices in anaphora resolutions, possessives and gender agreement.

PUBLISHER(S)

Google LLC

FUNDING

Google LLC

INDUSTRY TYPE

Corporate - Tech

FUNDING TYPE

Private Funding

DATASET AUTHORS

Anja Austermann, Google Michelle Linch, Google
Romina Stella, Google Kellie Webster, Google

DATASET CONTACT

translate-gender-challenge-sets@google.com

DATASET PURPOSE(S)

Testing

KEY APPLICATION(S)

Machine Translation, Gender Accuracy

ACCESS COST

Open Access

PRIMARY MOTIVATION(S)

Study gender accuracy in translations beyond the sentence in demographic and occupations diversity for fairness research.

INTENDED AND/OR SUITABLE USE CASE(S)

To evaluate gender accuracy on translations beyond the sentence (multiple sentences or passages). The set is focused on the presence of this specific linguistic phenomena to evaluate the most common contextual error:

- Pro-drop (Spanish ⟶ English)
- Neutral to gender-specific possessives (Spanish ⟶ English)
- Gender agreement (English ⟶ Spanish, German)

PRIMARY DATA TYPE(S)

Non-Sensitive Public Data about people

DATASET SNAPSHOT

Total Instances	138
Masculine biographies (entities)	63
Masculine biographies (countries)	51
Feminine biographies (entities)	63
Feminine biographies (countries)	57
Rock bands & sport teams (entities)	12
Rock bands & sport teams (countries)	12

SOURCE DATASET(S)

- Source text: English Wikipedia
- Target text: Professional translations

DESCRIPTION OF CONTENT

This dataset is based on publicly available data on public and/or historical figures (Wikipedia articles) at a given snapshot in time.

The dataset has 138 instances and each instance contains the first 8 to 15 sentences from a Wikipedia article. Articles are written in native English and have been professionally translated to Spanish and German. 126 of these instances represent a person with an associated stated gender and 12 are related with rock bands or sport teams (considered genderless).

PRIMARY DATA MODALITIE(S)	EXAMPLE OF ACTUAL DATA POINT*	HOW TO INTERPRET A DATAPOINT
Textual data	● **Source language:** en ● **Target Language:** de ● **Source text:** Kaisa-Leena Mäkäräinen (born 11 January 1983) is a Finnish former world-champion and 3-time world-cup-winning biathlete, who currently competes for Kontiolahden Urheilijat. ● **Translated text:** Kaisa-Leena Mäkäräinen (geboren am 11. Januar 1983) ist eine ehemalige Weltmeisterin und 3-malige Weltcup-Siegerin im Biathlon aus Finnland, die derzeit für Kontiolahden Urheilijat antritt. ● **Document ID:** 1 ● **Sentence ID:** 1-1 ● **Perceived gender associated with the entity:** Female ● **Name of the Entity:** Kaisa Mäkäräinen ● **Link to the original Wikipedia article:** https://en.wikipedia.org/wiki/Kaisa_M%C3%A4k%C3%A4r%C3%A4inen	**Each datapoint** refers to a central entity that can be a person (stated as feminine or masculine), a rock band or a sport team (considered genderless). **Each entity** is represented by a long text translation (multiple connected sentences or continuous passage referring to that main entity). Each datapoint will have the following data: ● **Source language:** Language of the original text ● **Target Language:** Language of the translation ● **Source text:** Text from Wikipedia in source language (special characters and quotes removed) ● **Translated text:** Translation of the Wikipedia source text into the target text ● **Document ID:** ID generated to identify all the sentences belonging to the same passage. ● **Sentence ID:** Composed by the Document ID and Sentence number in the passage. ● **Perceived gender associated with the entity:** identified as Female, Male, Neutral ● **Name of the Entity:** Name of the main entity according Wikipedia ● **Link to the original Wikipedia article:** Link to the Wikipedia article at the time of extraction. Please consider that content in Wikipedia articles can be modified so differences may be found if the article has been re-edited.

LICENSE TYPE(S)

CC-BY-SA 3.0

LICENSE BREAKDOWN

Source text has been extracted from English Wikipedia articles, which is made available under the CC-BY-SA 3.0 Unported license. All the rest is synthetic data.

LICENSE PERMISSIONS (CC-BY-SA 3.0)

- Share — copy and redistribute the material in any medium or format.
- Adapt — remix, transform, and build upon the material for any purpose, even commercially.
- Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made

STATUS

Limited Maintenance

DATASET VERSION

Version	1.0
Last Updated	06/2021
First Release	06/2021

FIRST EDITION

Original data was collected late in 2020 and translated at the beginning of 2021.

MAINTENANCE PLAN

No refreshes planned, but the dataset may be updated to incorporate feedback..

DATA COLLECTION METHOD(S)

Scraped

Independent Paid Professional(s)

DATA SOURCES BY COLLECTION METHOD(S)

- **Scraped:** English Wikipedia (source text)
- **Translation:** Independent paid professional human translations (target text)
- **Annotated:** Human added labels and metadata

DATA SELECTION CRITERIA - SCRAPING

- Grouped people from Wikipedia according to their occupation, profession, job and/or activity.
- Divided all these instances based on geographical diversity (optimizing for diversity at the country level), to mitigate the skew to Western-individuals (using regions from census.gov as a proxy of geographical diversity).
- Focused on having equal representation of feminine and masculine entities. Labeled each instance as "Female" or "Male" based on these biographies containing gender-indicative terms to refer to the person (like she, he, woman, son, father etc.)

Note: The set doesn't include non-binary individuals as we couldn't find enough instances to accurately reflect the community.

EXCLUDED DATA

- Quotes numbers from Wikipedia sentences were removed.
- Titles from the Wikipedia articles were excluded.
- Images were not considered. The dataset is just text.

SUMMARIES OF DATA COLLECTION METHOD

- **Scraping:** Sentences extracted from Wikipedia documents. (Source text)
- **Translations:** Source text has been professionally translated into the target language. For Spanish translations, guidance to focus on pronoun-drop sentences. (Target text)
- **Human added labels and metadata**: Added source and target languages, ids, entity names, links and perceived gender labels.

SAMPLING METHOD(S)

Stratified Sampling

SAMPLING BREAKDOWN

Total data sampled	2000 entities
Sample size	138

SAMPLING CRITERIA

- **Country diversity:** Entities that belong to countries that had at least 3 entities were discarded
- **Minimum text length:** 8 - 10 sentences
- **Occupational Activity:** subjects played an active role in the field of their occupation, and the

		wikipedia article pertains directly to their occupation - **Perceived gender:** inferred based on gender-indicative words in descriptions provided within the article - **Budgets:** within limits of budget available to project

HUMAN ATTRIBUTE(S)	PERCEIVED GENDER DISTRIBUTION		
Perceived gender Geography	**Perceived Masculine Biographies** Individual Instances 63 Country Coverage 51	**Perceived Feminine Biographies** Individual Instances 63 Country Coverage 57	**Genderless Articles (Rock bands & sports teams)** Individual Instances 12 Country Coverage 12

GEOGRAPHIC DISTRIBUTION				
Biographies * *organized by region for readability.	**Africa** 0.79% Cameroon 0.79% Central African Republic 0.79% Ethiopia 1.59% Ghana 1.59% Kenya 0.79% Liberia 0.79% Mauritania 0.79% Mauritius 0.79% Namibia 1.59% Nigeria 1.59% Senegal 0.79% South Africa 0.79% Tunisia 1.59% Uganda	0.79% Philippines 0.79% Singapore 0.79% South Korea 0.79% Sri Lanka 0.79% Thailand 1.59% Taiwan **Europe** 0.79% Armenia 0.79% Austria 0.79% Denmark 2.38% England 1.59% Finland 0.79% France 0.79% Georgia 1.59% Germany	**Latin America and the Caribbean** 0.79% Antigua and Barbuda 1.59% Argentina 0.79% Barbados 1.59% Brazil 0.79% Cayman Islands 1.59% Chile 0.79% Colombia 0.79% Cuba 0.79% Curaçao 0.79% Dominica 0.79% Dominican Republic	0.79% Iraq 2.38% Israel 0.79% Jordan 1.59% Lebanon 0.79% Morocco 1.59% Pakistan 1.59% Turkey **North America** 0.79% Bahamas 0.79% Belize 2.38% Canada 1.59% Jamaica 2.38% United States

	0.79% Zambia 0.79% Zimbabwe **Asia** 1.59% China 0.79% Hong Kong 2.38% India 0.79% Indonesia 2.38% Japan 0.79% Malaysia 0.79% Mongolia 0.79% Nepal	0.79% Hungary 0.79% Iceland 0.79% Ireland 0.79% Italy 0.79% Lithuania 0.79% Netherlands 0.79% Norway 1.59% Russia 0.79% Scotland 0.79% Spain 0.79% Sweden 0.79% Ukraine 0.79% Wales	0.79% Guatemala 0.79% Mexico 0.79% Paraguay 0.79% Trinidad & Tobago 0.79% Uruguay 0.79% Venezuela **Near East** 0.79% Algeria 0.79% Egypt 2.38% Iran	**Oceania** 2.38% Australia 0.79% Fiji 0.79% Micronesia 2.38% New Zealand 0.79% Palau 0.79% Papua New Guinea 0.79% Tonga 0.79% Tuvalu
GEOGRAPHIC DISTRIBUTION **Bands and sport teams**	**Africa** 8.33% Kenya 8.33% Nigeria 8.33% South Africa	**Asia** 8.33% Japan 8.33% South Korea 8.33% India	**Europe** 8.33% Spain 8.33% Sweden 8.33% Russia	**Latin America and the Caribbean** 8.33% Argentina 8.33% Brazil **Oceania** 8.33% Australia

LABELING METHOD(S)	LABEL TYPE(S)	LABELING PROCEDURE
Human labels Algorithmic labels	**Human labels** • Perceived gender: Annotated by raters based on gender-indicative words on the source text. **Algorithmic labels** • Document ID: generated by Google internal system.	**Human labels** • Perceived gender labels are based on the presence of gender-indicative terms in the article as explained in the "Data Selection Criteria" section. The label "neutral" was used for rock bands and sports teams. **Algorithmic labels**

	• Sentence ID: sequential number based on the location of the sentence in the paragraph • URL: extracted from Wikipedia • Entity name: extracted from Wikipedia.	• Entity Name was extracted from the title of the Wikipedia article. The URL redirects to the article version when the dataset was created. • Document IDs were assigned based on document ordering. Sentence IDs are based on the location of the sentence in the document.