




Manual Evaluation - Interpretability

The hereby presented manual evaluation task concerns determining whether the automatic translation of mGente sentences is NEUTRAL or GENDERED, and if neutralizations represent suitable translation for xai analyses.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
ID	SET	SRC	REF-G_ann	translation	gold_neutralit_y_label	Neutrality_type	Gendered_type	comments										
1341	Set-N	In addition, three things are needed: political will, fishermen who think like businessmen, and the interaction of consumers and entrepreneurs in the market place.	Además, tiene que existir una armonía triple: la voluntad política, <G>el<G> <G>pescador<G> que piensa de modo económico y la conjunción de <G>consumidores<G> y <G>empresarios<G> en el mercado.	Además, se necesitan tres cosas: voluntad política, personas que pescan y piensan como empresarios, y la interacción entre consumidores y empresarios en el mercado.														

Annotation Scheme Overview

Question	Condition	Label	Follow-up Questions
Can you analyse the translation of human entities?	<div>  No → (gibberish, incomplete, ungrammatical) </div> <div>  Yes → </div>	<div>NA</div> <div>—</div>	→ <i>Follow-up below</i>
<div>  What kind of translation is it? </div>	<div>Neutral</div> <div>Gendered</div> <div>Mixed</div>	<div>NEUTRAL</div> <div>GENDERED</div> <div>MIXED</div>	<div> What is the neutrality type? → LITERAL or INCLUSIVE, if normal GNT leave blank </div> <div> Does it use double forms? → If YES, mark double forms; if NO, leave blank </div> <div> Does it use double forms? → If YES, mark double forms; if NO, leave blank </div>

Guidelines

Given a source English sentence from mGente, evaluate if its automatic translation contains gendered forms (either masculine or feminine) or if it is a neutral translation.

For each sentence, indicate:

- **Gold_neutrality_label** (COLUMN P): pick one of the following labels:
 - **NEUTRAL:** The output sentence does not contain any gendered expressions to refer to human beings.
 - **GENDERED:** The output sentence contains only gendered expressions to refer to human beings.
 - 🖐️ double masculine/femine forms count as GENDERED.
 - **MIXED:** The output sentence preserves one or more gendered expressions to refer to human beings, but also one or more neutralizations.¹
 - **NA:** the output is gibberish and cannot be evaluated. Use this label only under the following conditions:
 - The output is overall very bad, not if it contains only some local errors
 - The gibberish/ungrammatical part of the output concerns the **translation of human entities** (either gendered or neutral) that are relevant to our task

🖐️ If you have doubts about what constitutes an acceptable Gender-Neutral Translation (GNT) in our task, look at the GNT guidelines used to create mGeNTE.

- **Neutrality_type** (COLUMN Q): This annotation layer only applies if the translation is NEUTRAL. You can pick two labels:
 - Indicate **LITERAL** if the NEUTRAL output translation does not seem a neutralization attempt, but rather looks like an incidental translation that, by being quite literal to the source, avoids any gender marking incidentally. Simply put, it is the most obvious translation the model might have produced, which by chance is also neutral.²
 - Use this label if the sentence contains only LITERAL neutralization types. E.g. This label should not be used if you find a correct GNT and one LITERAL neutralization.
 - Indicate **INCLUSIVE** if the output translation is inclusive using gender-fair strategies that differ from those expected in the GNT task (see our [GNT guidelines](#)). This might include e.g. the *-form for German, the schwa for Italian, the @ or “e” for Spanish.
 - Use this label if you find one or more INCLUSIVE type of neutralization

¹ This might also include e.g. the use of a gendered article + an inclusive noun, e.g. **Los** (*masc.*) *amigue* (**neomorpheme*)-- “The friends”

² This additional layer of information serves to get a sense of whether the model might have generated a neutral translation by “accident/chance” or if it seems to be able to solve the GNT task.

👏 If the neutralized output is neither literal nor inclusive, leave this empty.

- **Gendered_type** (COLUMN R): this annotation layer only applies if the translation is GENDERED OR MIXED.
 - Indicate **DOUBLE FORMS** if the output contains double feminine and masculine expressions, e.g. *Hello friends* ⇒ *es: Hola amigas/os, amigas y amigos*
- **Comment** (COLUMN S): Use this space to add optional comments that you deem relevant. E.g. the neutralization strategy used, or any trends that you might note.