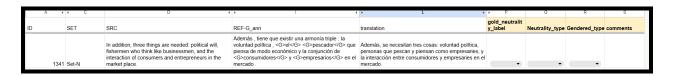
Manual Evaluation - Interpretability

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.



Annotation Scheme Overview

Question	Condition	Label	Follow-up Questions
Can you analyse the translation of human entities?	• •	NA	
	✓ Yes →	_	→ Follow-up below
✓ What kind of translation is it?	Neutral	NEUTRAL	What is the neutrality type? → LITERAL or INCLUSIVE, if normal GNT leave blank
	Gendered	GENDERED	Does it use double forms? → If YES, mark double forms; if NO, leave blank
	Mixed	MIXED	Does it use double forms? → If YES, mark double forms; if NO, leave blank

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 <u>does not contain any</u> gendered expressions to refer to human beings.
 - GENDERED: The output sentence <u>contains only</u> gendered expressions to refer to human beings.
 - double masculine/femine forms count as GENDERED.
 - MIXED: The output sentence <u>preserves one or more gendered expressions</u> to refer to human beings, but also <u>one or more neutralizations</u>.
 - NA: the output is gibberish and cannot be evaluated. Use this label <u>only under</u> the following conditions:
 - The output is overall very bad, <u>not</u> 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 <u>only applies if the translation is NEUTRAL.</u> 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 <u>only</u> 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 <u>gender-fair</u> strategies that differ from those expected in the GNT task (see our GNT <u>guidelines</u>). This might include e.g. the *-form for German, the schwa for Italian, the @ or "e" for Spanish.
 - Use this label if you find <u>one or more INCLUSIVE</u> type of neutralization

¹ This might also include e.g. the use of a gendered article + an inclusive noun, e.g. **Los** (*masc.*) amigu**e** (*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 <u>only applies if the translation is</u> <u>GENDERED OR MIXED.</u>
 - 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 <u>optional</u> comments that you deem relevant. E.g. the neutralization strategy used, or any trends that you might note.