Gender-inclusive language and machine translation: from Spanish into Italian

Antonella Bove
Università Ca' Foscari Venezia
antonella.bove@unive.it

Abstract

Direct gender-inclusive language strategies represent a significant challenge for automatic translation systems because they often involve non-standard forms that systems are not trained to recognize or replicate accurately. This paper aims to shed light on the way in which four artificial intelligence systems interact with Spanish inclusive strategies in their translation into Italian through case study analysis within an augmented translation perspective (Kornacki & Pietrzak 2025). For this purpose, authentic academic texts were used, which can therefore constitute real translation assignments. The outputs of four AI systems were compared and analysed: the neural systems of DeepL and Google Translate and the generative systems of ChatGPT and Gemini.

1 Introduction

Direct gender-inclusive language is a discursive practice that introduces the use of new forms and strategies to make women and different non-binary gender identities more visible (Román Irizarry et al. 2025). Spanish uses split forms and gender doublets (*los niños y las niñas*, *los/as candidatos/as*), the neomorpheme *-e*, and typographic signs such as @ and *x* (Román 2025, Papadopoulos 2022). Similarly, Italian employs split forms and gender doublets (*i bambini e le bambine*, *i/le candidati/e*), the schwa (*ə*) as a neomorpheme, and the asterisk (*) as a typographical symbol. While gender doublets and the @ sign aim to make women visible within a binary framework, the others are intended to give visibility to non-binary gender identities as well (Escandell-Vidal 2020, Giusti 2022). In the Augmented Translation era, it is crucial to study and evaluate the performance of automatic translation systems to determine whether and to what extent diverse gender representations are maintained in the process of translation (Gutt 2000, López 2021).

2 Methodology

The methodology involved three main steps: data collection, annotation, and analysis.

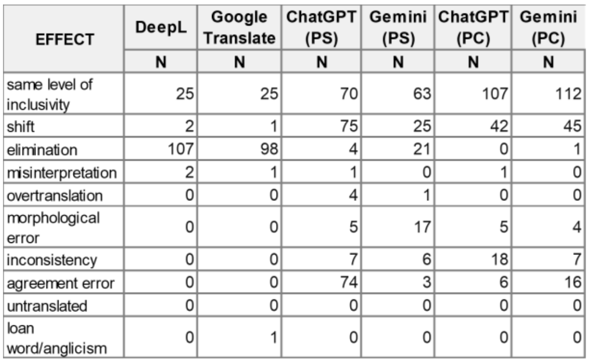
Academic texts originally written in Spanish were gathered; from those specific segments were extracted. Segment-level analysis allowed for the creation of a more diverse corpus. In total, approximately 20 instances were collected for each inclusive language strategy examined: split forms, doublets, the neomorpheme *-e*, the sign @ and the letter *x*. These segments were then translated using four artificial intelligence systems: two neural translation systems (*DeepL* and *Google Translate*) and two generative AI systems (*ChatGPT* and *Gemini*). Additionally, two prompts were created for the last two systems: a complex one (PC) and a simple one (PS). Based on the few-shot prompting technique, the former (PC) included information about the translation task, namely (i) source and target languages, (ii) intended purpose (publication) and (iii) examples to guide the system. Specifically, it provided examples linking source strategies with corresponding target strategies. The latter (PS) included the same information, except for the examples.

The translated outputs were annotated using UAMCorpusTool (O'Donnell 2008), and a taxonomy of effects was created. This taxonomy categorizes translation outcomes into several effects such as: *same level of inclusivity*, with further subdivisions that indicate whether this was achieved through a (non)standard linguistic strategy or because the equivalent in the TL is neutral or a common-gender noun; *overtranslation*, which marks all outputs that contain unnecessary inclusive language marks;

misinterpretation refers to erroneous interpretation of the machine with a consequent meaning error; *elimination* of inclusive markers that were used in the source text (but without omitting the word itself); *shift* in the level of inclusivity, from binary to non-binary and vice versa, changing representations arbitrarily; *morphological error* concerns the target text, as it refers to cases in which the system uses a gender-inclusive mark without respecting the morphological constraints of the target language; *inconsistency* refers to the use of different gender-inclusive marks within the same segment; *untranslated* label refers to untranslated words; *agreement error* is for cases in which the system uses a gender-inclusive mark that does not agree with the other words in the phrase; *loan word/anglicism* concerns the translation of a word with a loan word from a language without grammatical gender, such as English, making it more inclusive.

3 Results

The analysis revealed significant differences between the generative and non-generative systems.



EFFECT	DeepL	Google Translate	ChatGPT (PS)	Gemini (PS)	ChatGPT (PC)	Gemini (PC)
	N	N	N	N	N	N
same level of inclusivity	25	25	70	63	107	112
shift	2	1	75	25	42	45
elimination	107	98	4	21	0	1
misinterpretation	2	1	1	0	1	0
overtranslation	0	0	4	1	0	0
morphological error	0	0	5	17	5	4
inconsistency	0	0	7	6	18	7
agreement error	0	0	74	3	6	16
untranslated	0	0	0	0	0	0
loan word/anglicism	0	1	0	0	0	0

DeepL and *Google Translate* tended to eliminate gender inclusive marks, including split forms and doublets, defaulting to masculine forms. On the contrary, *ChatGPT* and *Gemini* demonstrated a better capacity to maintain inclusivity but did so inconsistently. Moreover, these systems generated syntactical and morphological errors in some cases, likely due to inadequate training on direct gender inclusive language forms. Additionally, there were numerous instances of arbitrary shifts in the gender marking, from binary to non-binary and vice versa, significantly altering the represented identities. Finally, few-shot prompting resulted in better outputs compared to basic prompting, suggesting an interaction between prompt accuracy and translation accuracy.

4 Conclusion & Future Work

The study concludes that generative AI systems show greater potential for the translation of gender-inclusive texts when guided by well-designed prompts, compared to neural machine translation systems that are not effective for this purpose as they tend to eliminate almost all the direct strategies. However, two key methodological limitations should be acknowledged, due to limited resources and the pioneer nature of the study: a) the annotation was conducted by a single annotator, preventing the calculation of the inter-annotator agreement and therefore limiting empirical validation and b) the analysis was carried out on a segment level, which, while useful for isolating specific linguistic phenomena, does not account for broader discourse-level or co-textual influences that could affect translation outputs. Future research should involve multiple annotators and adopt a document-level approach to enhance the reliability and generalizability of the findings.

References

Escandell-Vidal, M. V. (2020). *En torno al género inclusivo*. IgualdadES, 2, 223–249.

Giusti, G. (2022). *Inclusività nella lingua italiana: come e perché. Fondamenti teorici e proposte operative*. DEP. Deportate, Esuli, Profughe, 48, 1–19.

Gutt, E.-A. (2000). *Translation and relevance: Cognition and context* (2nd ed.). Taylor and Francis.

Kornacki, M., & Pietrzak, P. (2025). *Hybrid workflows in translation: Integrating GenAI into translator training*. Routledge.

López, Á. (2021). *Cuando el lenguaje excluye: Consideraciones sobre el lenguaje no binario indirecto*. Open Science Framework.

O'Donnell, M. (2008). *The UAM CorpusTool: Software for corpus annotation and exploration*. In Proceedings of the XXVI Congreso de AESLA, Almeria, Spain, 3–5 April 2008.

Papadopoulos, B. (2022). Una breve historia del español no binario. Deportate, esuli, profughe, 48, 31-39.

Román Irizarry, A., Beatty-Martínez, A. L., Torres, J., & Kroll, J. F. (2025). *"Todes" and "Todxs", linguistic innovations or grammatical gender violations?* Cognition, 257, 106061.