Assessing the Performance of Large Language Models for Bilingual Term Extraction in Interpreting and Translation

Mahmoud Gaber
mahmoudgaber@uma.es
IUITLM, University of Malaga
21/05/2025, CIUTI Conference 2025

Table of contents

01

Introduction

A Brief Overview of the Study

Introduction

Large Language Models provide us with efficient tools for various Natural Language Processing tasks. This research evaluates four AI tools—ChatGPT, DeepSeek, Copilot, Gemini and Manus—for extracting Spanish-Arabic bilingual terminology for translation and interpreting purposes. Given the lack of research on non-Indo-European languages, the study assesses each tool using Precision, Recall, F-score, and Accuracy to determine their suitability for professional use and improve AI-driven terminology management.

02

Rationale

Rationale and Motivation

Justification and Motivation

- *Existing research on **AI-based terminology extraction focuses mainly on European language** pairs.*
- *As Large Language Models (LLMs) are increasingly used in training and professional settings, **their reliability in underrepresented language pairs must be assessed**.*
- ***Bilingual terminology is essential for interpreting and translation**, particularly in specialised domains (Veisbergs 2006; Rodríguez and Schnell 2009; Corpas Pastor and Fern 2016; Cavallo 2017; Corpas Pastor 2018; Sales 2024).*
- *Other methods for bilingual terminology extraction, such as **corpus-based approaches**, present several **challenges**, including:*

- ***alignment issues*** (e.g., Castillo Rodríguez 2011; El-Farahaty, Khallaf, and Alonayzan 2023, Gaber 2025);
- ***difficulty to obtain parallel corpora*** *(Daille 2012; Delpech et al. 2012; Hazem and Morin 2016b; Kontonatsios 2015).*

03

Objectives

Research objectives

Research Objectives

- Evaluate and compare the performance of ChatGPT-4o, DeepSeek, Copilot, Gemini and Manus
- Determine the most suitable AI tool for Spanish-Arabic bilingual term extraction
- Identify areas for improvement in AI-driven terminology management

04

RQs

Research Questions

Research Questions

- How effectively do **LLMs extract bilingual terminology** for Spanish-Arabic translation and interpreting?
- Which AI tool performs best in terms of **Precision, Recall, F-score, and Accuracy**?
- What are the **limitations** and **potential enhancements** in AI-assisted terminology extraction?

05

Methodology

Methodological Approach

Methodology

- **Dataset** and Corpus Selection:

Domain	:	Medical (Ophthalmology)
Corpus **type**	:	Comparable Corpus
Languages	:	(Spanish-Arabic)
Corpus **Size**	:	Spanish (1.812) – Arabic (8.023)

Gold set (human-curated Gold Set of 75 terms) in Spanish

- **AI tools** Selected: **ChatGPT**-4o, **DeepSeek**-R1, **Copilot, Gemini and Manus**
- **Prompting** Techniques: mixed approach (CREATE and CO-STAR)
  - **Terminology extraction**
  - **Equivalents extraction and suggestion**

Methodology

- **Evaluation Metrics**

Compare **5 term lists** (extracted by 5 different LLMs) against a **human-curated Gold Set** and compute:

Precision = [Correctly Extracted Terms] / [Total Terms Extracted by LLM]
Recall = [Correctly Extracted Terms] / [Total Terms in Gold Set]
F-score = 2 * (Precision * Recall) / (Precision + Recall)
Accuracy = [Correct Terms] / [Total Terms in Gold Set + Incorrect LLM Terms]
Jaccard Index = [Gold Set ∩ LLM Terms] / [Gold Set ∪ LLM Terms]

06

Rs & D

Results and Discussion

Performance Analysis Results

Model Name	Precision (%)	Recall (%)	F-score (%)	Accuracy (%)	Jaccard Index (%)
ChatGPT-4o	84.44	50.67	63.33	46.34	46.34
Copilot	68.89	41.33	51.67	34.83	34.83
DeepSeek-R1	87.50	46.67	60.87	43.75	43.75
Gemini	45.26	57.33	50.59	33.86	33.86
Manus	54.88	60.00	57.32	40.18	40.18

Conclusion-I

A clear **trade-off between precision and recall** emerged across models. Models with **higher precision** (DeepSeek, ChatGPT) tended to have **lower recall**, while models with higher recall (Manus, Gemini) demonstrated lower precision.

Conservative extractors: DeepSeek and ChatGPT appear to be more selective, prioritizing correctness over comprehensiveness.
Liberal extractors: Manus and Gemini seem to cast a wider net, capturing more gold standard terms but including more incorrect terms.

Conclusion-II

This evaluation reveals significant variations in how different LLMs approach terminology extraction in specialized domains. While ***DeepSeek and ChatGPT demonstrated superior precision, and Manus showed stronger recall, no model achieved excellent performance across all metrics****. The* ***highest F-score (ChatGPT's 63.33%) still falls below*** *what would be considered strong performance in most applications*

Conclusion-III

These findings suggest that **current LLM-based terminology extraction systems have substantial room for improvement**, particularly in achieving better balance between precision and recall.

Common Issues in Terminology Extraction Using Large Language Models (LLMs)

- Limited identification of **multi-word terms**.
- Difficulty recognising compound terms whose components do not appear together (e.g., *catarata* may be congenital or acquired).
- Inconsistent recognition and extraction of **acronyms**.
- Redundancy and **repetition** of extracted terms.
- Inclusion of items that do **not** qualify as **valid terms**.

Thanks!

Q&A + Open Discussion

mahmoudgaber@uma.es

References

Castillo Rodríguez, Cristina. 2011. “La Alineación de un Corpus Paralelo Multilingüe: Propuesta de Fases para la Didáctica de Traducción Especializada Inversa.” *Cadernos de Tradução* 27 (1): 117–142. https://doi.org/10.5007/2175-7968.2011v1n27p117

Corpas Pastor, Gloria, and Lily May Fern. 2016. "A Survey of Interpreters' Needs and Practices Related to Language Technology." Technical Paper FFI2012-38881-MINECO/TI-DT-2016-1.

Jaccard, Paul (1901). "Étude comparative de la distribution florale dans une portion des Alpes et des Jura". *Bulletin de la Société vaudoise des sciences naturelles (in French). 37(142): 547–579.*

Sales, Dora. 2024. “Professional Translators’ and Interpreters’ Views on Information Competence: An Exploratory Qualitative Study from the Spanish Context.” Journal of Librarianship and Information Science 56 (3): 743–59. https://doi.org/10.1177/096100062311641

Veisbergs, Andrejs. 2006. “Dictionaries and Interpreters." EURALEX Proceedings. 1219–1224. https://www.euralex.org/elx_proceedings/Euralex2006/146_2006_V2_Andrejs%20VEISBERGS_Dictionaries%20and%20Interpreters.pdf

Acknowledgment:

This work was carried out in the framework of the following research projects: Postdoctoral research contract (PPIT-UMA), PIE22-135 (2022/23-2023/24), VIP II (PID2020-112818GB-I00/AEI/10.13039/501100011033), RECOVER (ProyExcel_00540), DIFARMA (HUM106-G-FEDER), and DÍGAME (JA.A1.3-06).