Characterization of Institutional Texts for an Automated Golden Standard: Enhancing Machine Translation Quality Assessment between English and Spanish

María Luisa Romana García[1][0000-0002-9962-1629] and Blanca Hernández Pardo[2][0000-0001-9005-7577]

[1] Universidad Pontificia Comillas, C. Universidad Comillas, 28108 Madrid, Spain
mlromana@comillas.edu
[2] Universidad Pontificia Comillas, C. Universidad Comillas, 28108 Madrid, Spain
bhpardo@comillas.edu

Abstract. The purpose of this paper is to collect a set of features that can contribute to the linguistic characterization of the institutional textual genre. The aim is to describe as exhaustively as possible the archetypal text to be obtained as a target text in this type of specialized translation. The tools used were Orange Data Mining© and Google Colab (Python code), and the data was obtained using the following processing mechanisms: word cloud, text preprocessing (cleaning, tokenization, normalization, lemmatization and PoS annotation). With these tools, lexical and grammatical frequencies, lexical and documentary embeddings, cosine distances, hierarchical clustering, and 20-component dimensionality reduction (t-SNE) were extracted.
As a result, a series of useful descriptive parameters have been obtained for the characterization of model texts for economic translation of institutional domains into Spain Spanish: lexical and terminological density, phraseological and terminological lexicalizations, grammatical frequencies, and semantic maps. In conclusion, the study provides several quantifiable features that characterize the analyzed register and opens the way for further research to deepen these parameters and develop the research by searching for complementary parameters until a complete and exhaustive picture of the reference model in this genre is obtained.

Keywords: Machine Translation, Golden Standard, Translation Quality Assessment, Specialized Translation, AI Processing.

1 Introduction

Machine Translation (MT) has greatly enhanced translation accessibility, sparking frequent concerns about its potential to replace human translators (Amini et al. 2024 p. 743). However, professional translation, especially for critical documents such as legal agreements or financial statements, demands accuracy and human oversight, where quality assurance is indispensable (Moorkens and Gerberof 2024). Despite the advancements, MT is still plagued by reliability issues and inherent limitations (Lyu et al. 2024a; He et al. 2022), prompting a shift towards Large Linguistic Models (Lyu et al.

New Trends in Translation and Technology, pages 138–155
3–6 July, 2024. ©NeTTT 2024

2024b). Documents with significant responsibility require a combination of MT and human expertise to ensure accuracy and reliability. As Artificial Intelligence (AI) continues to permeate professional realms, the human component remains crucial for complex and socially relevant tasks. Alongside human expertise, a robust Quality Assessment (QA) mechanism is essential to evaluate raw MT outputs. Traditionally, translation quality has been assessed against a "Golden Standard" (GS), a human translation used for comparison. This process involves understanding the text type and genre, especially in specialized translations that handle legally, politically, economically, or socially significant texts.

The concept of "fit-to-purpose" or Error-Tolerant Translation (ETT) is important, emphasizing communication goals despite potential errors, thus necessitating human involvement for accuracy. Error tolerance significantly impacts the required human effort and associated costs. Despite AI's efficiency, human oversight is crucial for precise and reliable translations (Gil Sanromán et al. 2021).

Our research aims to develop an abstract automated GS by analyzing institutional texts to assess (a) the error tolerance of a Source Text (ST) and (b) the quality of a Target Text (TT). By comparing the GS with the ST, we can determine the document's error tolerance level. As a first step —and the research goal at this stage—, we are exploring the possibilities of characterizing texts which are not the product of translation but institutional Spanish texts published by private and public entities in the economic and financial field. After that, we will need to do the same with a corpus of translated texts in order to be able to compare both characterizations and see whether we can eventually arrive at an abstract GS usable by the industry for automation purposes. Accordingly, the primary focus of this paper is to contribute to the characterization of institutional native texts. This involves a comprehensive analysis to identify key parameters able to influence the assessment of TT quality.

This paper is structured as follows: the introduction provides an overview of the study's objectives and significance; the literature review examines previous research on translation quality assessment (TQA) and methodologies employed in this field; the methodology section details the steps taken to analyze institutional texts using AI techniques; the results section presents the findings of the study; the discussion explores the implications of these findings for TQA and MT evaluation as well as the limitations of this work; and the conclusion outlines the contributions of this research and potential directions for future work.

2 Literature review

This literature review examines key themes in translation quality assessment (TQA) relevant to our research. It focuses on genre adaptation, evaluation metrics, and specialized translation, laying the groundwork for developing an automated Golden Standard (GS) for assessing English-Spanish institutional translations.

2.1 Genre Adaptation and Cultural Context in Translation

Adapting a Target Text (TT) to its target culture is crucial in translation studies, balancing literal and non-literal approaches. Deviations between the TT and the ST are often considered errors (Hedayati & Yazdani, 2000). House's TQA model (1997a, 1997b) evaluates translations based on language register and genre, categorizing errors as overt (linguistic/textual) or covert (register/genre). Muzii (2006) highlights the impact of technological, cultural, and market changes on translation standards, stressing economic viability and customer satisfaction. Evaluating translation quality, especially for institutional and legal texts in Spanish, requires understanding the textual characteristics of the genre within the target culture. This involves identifying textual features that align with the communicative goals of the ST. Extensive literature on QA emphasizes empirical data and professional collaboration (Eyckmans et al., 2012), user-defined error weights (Daems et al., 2013), and structured evaluation (Huertas Barros et al., 2015). Rating systems like MQM and DQF (Babych, 2014) address methodological inconsistencies. Recent advancements include COBTAS, an automated educational platform (Akrami et al., 2018), and CPIE for enhancing assessment objectivity (Akbari and Shahnazari, 2019). Comprehensive quality frameworks such as MQM and DQF (Mariana, 2014; Lommel, 2018) classify errors to set quality metrics. The MAP tool supports a mixed qualitative-quantitative approach to assess quality (Martínez Mateo et al., 2016), while Portilho & Drugan (2018) review quality measurement methods, comparing industry and academic approaches. Ebeling and Ebeling (2020) highlight the influence of corpus type on cross-linguistic equivalence, recommending both monolingual and translated texts for analysis.

Authors agree on the difficulty of creating an automatic model for evaluating translation quality due to the variety of possible errors and solutions in any given TT. Typically, a single human text is used as a benchmark for AI models.

2.2 Evaluation Metrics and the Construction of a "Model Text"

Evaluating machine translation (MT) quality has evolved, with traditional methods comparing translated texts to a human-created Golden Standard (GS) using metrics like Levenshtein distance (Finch et al. 2004; Hernández Pardo & Romana García, 2021; Volk & Harder 2007; Wichmann et al. 2010). However, this approach has limitations, leading to the exploration of alternative techniques. Fan (1990) introduced fuzzy subset theory for nuanced evaluation metrics, while Tajvidi (2005) proposed a scoring system emphasizing human judgment. Banchs et al. (2011) found strong correlations between human and machine evaluations, focusing on appropriateness and fluency. Recent research by Han (2020) and Han et al. (2021) suggests including cross-cultural consistency tests in QA models. This proposal is particularly relevant as it acknowledges the dynamic nature of language and the need for evaluation methods that account for cultural nuances. Rivera Trigueros (2022) explores the dominance of neural network translation models, such as Google Translate, and the use of both machine and human evaluation methods, often utilizing error taxonomies to improve MT systems. Advanced neural frameworks like COMET (Rei et al. 2020) represent significant

advancements in TQA, using metrics like BLEU, ROUGE, and METEOR (Cer et al. 2010; Nema et al. 2018; Zhang et al. 2024). Jaszczolt (2003) and Kurteš (2009) emphasize the importance of cultural and contextual factors in meaning interpretation, crucial for establishing a reliable GS. However, these tools face methodological and practical challenges, highlighting the need for a standardized QA framework (Giovannotti, 2023). The evaluation of translation adequacy involves comparing the TT to an ideal model or GS, often conceptualized as either an abstract notion or a carefully revised human-made text. Jaszczolt (2003) emphasizes the impact of cultural and contextual factors on meaning interpretation, while Kurteš (2009) employs corpus-based studies for contrastive analysis, demonstrating how linguistic features serve as benchmarks for TT evaluation. Rabadán et al. (2009) introduce the ACTRES methodology, which focuses on identifying translation universals to highlight low-quality translations.

The concept of "tertium comparationis", rooted in Renaissance debates on rhetoric and dialectics (Zhu, 2017), represents an ideal of textual perfection within a cultural context. Chen (2017) extends this concept by adding a semiotic layer to bridge linguistic gaps across language families. Hatim (2024) and Borja et al. (2014) emphasize the influence of linguistic and cultural factors on translation, particularly within technical genres. Faber Benítez (2009) explores the cognitive turn in translation studies, focusing on semantic and conceptual frameworks, which are crucial for understanding how meaning is constructed and conveyed in translation.

In summary, the literature highlights a range of methodologies and frameworks for assessing MT quality. These approaches underscore the complexity of translation assessment and the need for comprehensive models that account for linguistic accuracy, cultural context, and technological advancements. This paper aims to build on these foundations by establishing a detailed list of features and levels for evaluating AI-generated translations, ultimately contributing to the development of a more standardized and reliable QA framework.

2.3 Genres and specialized translation

Specialized translation, particularly in institutional texts, requires understanding genre-specific characteristics. Valero Garcés (1996) performs a detailed comparative analysis of economic texts in Spanish and American English, finding significant cross-cultural differences in rhetorical styles. English texts are more reader-oriented, using explicit rhetoric, while Spanish texts prioritize propositional content and an impersonal tone, often including more implicit content. This underscores the need for developing genre-specific evaluation metrics. Research by Sevilla Muñoz (1997), Zarco Tejada (1998), Portolés Lázaro (2002), Lorenzi Zanoletty (2005), Limón Aguirre (2013), Hennecke (2015), and others has explored phraseology, syntax, style, discourse, and intercultural elements through theoretical analysis. Fernández Parra (2007) categorizes repetitive language in specialized translations by function and structure, useful for both automatic and human translation. Studies on MT formulaic texts show varying progress across language pairs, highlighting the importance of genre-specific features (Gajer 2009; Faber & Ureña Gómez-Moreno 2012; Forchini & Murphy 2008; Hmida et al. 2016). Forchini & Murphy (2008) analyzed prepositional 4-grams in English/Italian texts,

finding higher prevalence of certain engrams, useful for building a GS. Subhi Khalil (2020) categorized complex nominals in economic discourse, while Romana García (2009) compared syntactic structures in economic texts translated from English to Spanish.

These studies collectively underscore the importance of genre-specific characteristics in translation quality assessment and provide a robust foundation for developing a GS for evaluating institutional texts. The literature highlights the complexity of translation assessment and the need for comprehensive models that account for linguistic accuracy, cultural context, and technological advancements. Our research aims to build on these foundations, establishing detailed features for evaluating AI-generated translations, contributing to a standardized and reliable QA framework for institutional texts.

3 Methodology

As stated, our research focuses on economic texts originally written in Spanish and published by Spanish private entities and public bodies. Our goal is to use a set of AI techniques to structurally define a sub-genre of monolingual institutional texts, identifying patterns or features for evaluating TTs without depending on a human-made GS. This research concentrates on the pair English/Spanish and a specialized sub-genre, aiming to create specific metrics to allow for a quantitative assessment.

All the files used for this study are included in a GitHub repository, where the text corpus, tables and lists of frequencies mentioned throughout this paper can be found: https://github.com/traia24/categ_esp.git.

3.1 Corpus

Corpus-based approaches enable the extraction of linguistic patterns and features that can be quantified and analyzed (Olohan 2004). By applying these principles, we aim to develop metrics that specifically measure the structural and linguistic characteristics of translated economic texts, providing a robust framework for quantitative evaluation.

We analyzed Spanish 45 PDF texts from three authoritative sources in Spain (private and public entities), published from May and June 13, 2023: the Banco de España (Bank of Spain), the Comisión Nacional del Mercado de Valores (CNMV, National Securities Market Commission), and the Círculo de Empresarios (Business Circle). Full texts can obtained from our GitHub repository: https://github.com/traia24/categ_esp.git. For the GS, we have selected three features: lexical frequencies, grammatical frequencies, and syntactic structure frequencies.

Table 1. Corpus used (summary of data).

Public / Private Body	Texts	# pages	# words
Banco de España	Annual Report	513	255,999
CNMV	Legislation, regulations and recommendations	512	206,665
Círculo de Empresarios	Macroeconomic analysis	533	150,764
	TOTAL analyzed	**1,558**	**613,428**

3.2 IA Techniques

We used the commercial app Orange Data Mining® (ODM)[1] to conduct most of the research AI tasks. ODM is a comprehensive open-source data visualization and analysis tool that leverages a visual programming approach to facilitate data analysis through a workflow-based interface. This interface is built around "widgets," which are modular components that perform specific functions within a data analysis pipeline.

We calculated cosine distances to measure the similarity between documents and applied hierarchical clustering to group similar documents. To visualize the high-dimensional data, we employed a 20-component dimensionality reduction technique using t-SNE. Data were obtained by the following processing steps:

Firstly, we used the Import Documents widget within ODM for corpus creation. This allowed us to gather and manage the documents needed for our study. Following this, we cleaned the corpus by removing stopwords using a custom file designed specifically for this purpose through the “Preprocess Text” widget. This step ensured that our text data was free from common, non-informative words that could skew the analysis. Next, we extracted word clouds using the respective widget in ODM (see Figure 1). This process involved iterating through a subloop where we continuously updated the stopwords file based on the word cloud results and then re-cleaned the text data with the “Preprocess Text” widget. This iterative refinement helped to improve the quality of our stopwords list and, consequently, the clarity of the word cloud. We proceeded with a comprehensive text pre-processing: cleaning, tokenization, standardization, and lemmatization of the text data. Part of this pre-processing also involved Part-of-Speech (PoS) tagging, which helped us to categorize words based on their grammatical functions. For data collection, we extracted various metrics such as word clouds, lexical frequencies, and grammatical frequencies (PoS) (see Figure 2). Additionally, document embedding techniques were used to convert text data into numerical form, enabling further analysis.

A flowchart detailing the entire step-by-step procedure can be obtained from the GitHub repository (“Procedure flowchart”).

Each of these steps in Orange Data Mining, facilitated by the appropriate widgets, contributed to a thorough and efficient analysis pipeline, allowing us to derive meaningful insights from our textual data.

4 Results

4.1 Word Cloud

We generated a word cloud for the entire corpus to visualize the most frequent lexical units. The word cloud revealed key terms that characterize the genre of institutional texts (Figure 1).

[1] https://orangedatamining.com/. See https://orangedatamining.com/docs/ for technical reference on widgets and technical procedures.

Fig. 1. Word cloud (complete corpus).

4.2 Lexical Frequencies

We calculated the frequency of each token in the corpus of institutional texts to identify the most frequent lexical units and Terminological Units (TUs). The complete list of these frequencies is included in the GitHub repository ("Full word frequencies"). The most frequent terms were "*empresa*" (11,760 occurrences), "*público*" (10,360 occurrences), and "*inversión*" (9,612 occurrences). As we will discuss, these are not terminological but phraseological units, which sets a severe limitation on the effectiveness of this method.

4.3 PoS Frequencies

We extracted the frequency of various parts of speech, focusing on nouns, adjectives, adverbs, and verbs. The results are visualized in Figure 2, with the detailed table provided in the GitHub repository ("Top 20 most frequent words").

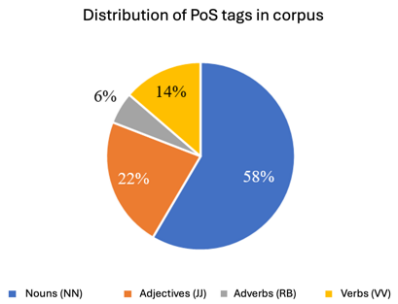

Fig. 2. Frequency of grammatical categories.

Nouns and adjectives together accounted for 80% of the total parts of speech, with nouns at 58% and adjectives at 22%. Verbs constituted 14%, and adverbs were the least common at 6%.

4.4 Extraction of Functional and Technolectal Lexicalization

We analyzed lexicalizations, which are words that frequently co-occur and form potential TUs. Our analysis identified the top 20 words with more than 6,000 occurrences, including "*empresa*" (11.760), "*público*" (10.360) and "*inversión*" (9612. Frequent lexical combinations such as "financial market policy" and "public sector" were also noted (Figure 3).

Fig. 3. Lexical combinations: "*Política del mercado financiero*" (financial market policy), "*sector público*" (public sector).

Having examined the 20 most frequent words, we apply the same steps to all relevant words in the corpus, i.e., all words in the word cloud (sorted by number of words in the lexical unit). This procedure is not very effective in confirming the usual expectations concerning the terminological density of a specialized text (Romana, 2012). Despite identifying frequent terms, our tools were less effective in confirming compound TUs, indicating the need for more sophisticated extraction methods in future research.

4.5 Lexical-semantic associations

Using hierarchical clustering, we grouped words according to their semantic relations, generating a table of general paradigmatic relations, provided in the GitHub repository ("Paradigmatic relations at the lexical-semantic level"). Dimensionality reduction via t-SNE produced a semantic map of the corpus, revealing two main regions (Figure 4).

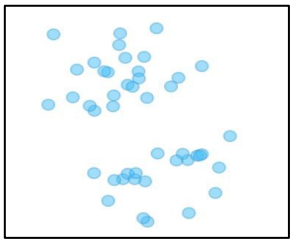

Fig. 4. Semantic map from t-SNE analysis (MDG)

The two regions identified were «Zone 1» (regulatory compliance issues in financial institutions) and «Zone 2» (public policy topics: Spain, the EU, public spending and economic policies). The semantic analysis thus confirms the validity of the corpus.

4.6 Semantic Clustering by Lexical Indicators

The semantic analysis carried out on the MDGs (t-SNE) shows the semantic distribution of the ideas common to the two regions. Thus, the upper domain corresponds to regulatory compliance problems in private entities (and public entities in the background). Figures 5 and 6 illustrate the semantic clustering of texts.

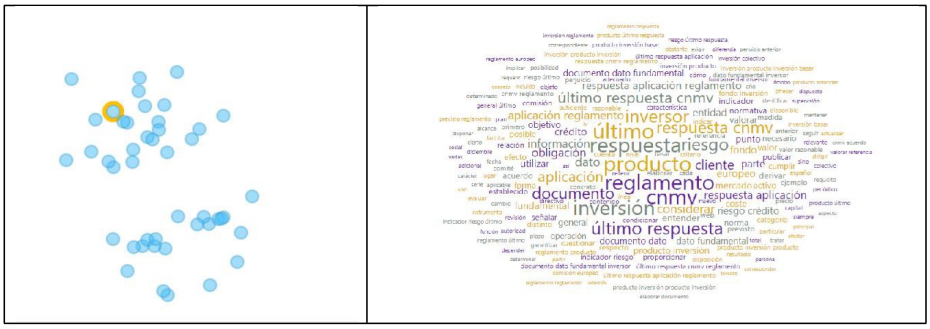

Fig. 5. Grouping of texts according to semantic factors by documentary embedding: Zone 1

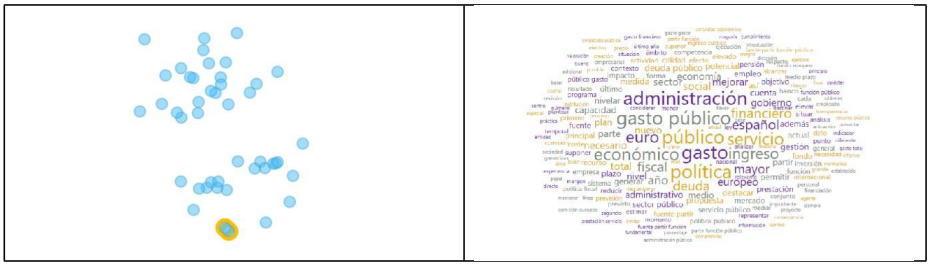

Fig. 6. Grouping of texts by semantic factors by documentary embedding: example Zone 2

Zone 1 clusters around terms like "product," "response," "regulation," and "CNMV," indicating texts focused on financial regulatory compliance. Zone 2 includes terms like "Spain," "European Union," and "public spending," highlighting texts on public economic policies.

4.7 General Semantic Characterization

Thematic patterns were observed in the unabridged areas, providing a broader understanding of the semantic structure (Figure 7).

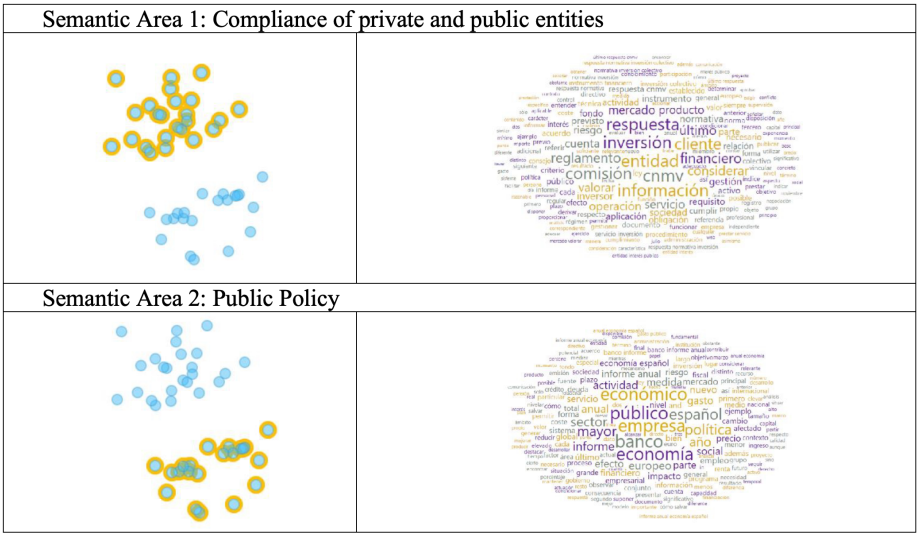

Fig. 7. General Semantic Characterization (Zones 1 and 2): Thematic Patterns

5 Discussion

The results indicate that institutional texts can be characterized by specific lexical and grammatical features, confirming expectations from the literature. Key findings include:

- High lexical density of specialized language.
- Dominance of nominal categories (nouns and adjectives).
- Identification of frequent lexical combinations and potential TUs.

The primary aim of this study was to explore the first steps towards the creation of automated Golden Standard (GS) for evaluating the quality of institutional translations from English to Spanish. The findings from our analysis provide significant insights into the linguistic and structural characteristics of institutional texts, which are essential for creating a reliable GS.

5.1 Lexical and Part-of-Speech (PoS) Frequencies

The high frequency of specific lexical units such as "*empresa*," "*público*," and "*inversión*" highlights the focus areas within institutional texts. These words are of a phraseological rather than terminological nature (Lorente, 2000). This confirms that typical macroeconomic texts show a high density of specialized language, but the lexical level seems to be restrained to the phraseological domain, which would not be as useful in assessing MT quality.

Out of the 98 most common words in the corpus, the first TU is "*inversión*", ranking third with 9612 occurrences. 34 of these 98 words are also listed in the reference BBVA

financial lexicon,[2] confirming that lexical density is a key feature of this text type. Marking the TU, we find a minimum density of 22% among the 98 most frequent words (Figure 8). However, this AI technique seems to merge terminological and phraseological, which would not render a successful metric in that these two spheres need to be clearly distinguished in a specialized translation. Therefore, more research needs to be done in this regard.

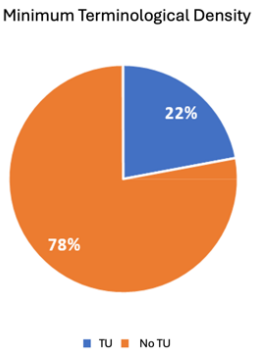

Fig. 8. Minimum terminological density of the corpus (only the first 98 lexical frequencies are measured).

5.2 Functional and Technolectal Lexicalization

The identification of frequent lexical combinations, such as "financial market policy" and "public sector," indicates the presence of standardized phrases within institutional texts. However, the difficulty in extracting compound Terminological Units (TUs) suggests that current tools may need enhancement to accurately capture these nuances. Future development of the GS should include advanced methods for detecting and evaluating such compound TUs to ensure comprehensive quality assessment

5.3 Parts of Speech

As for the PoS level, our results confirm an already known characteristic (Álvarez García, 2011, p. 287; Vangehuchten, 2004, p. 1134): 58% nouns and 22% adjectives, the figures do confirm that nominal categories represent 80% of the total categories, clearly confirming the literature observations. The remaining 20% are mainly verbs (14%), with adverbs being the least common category. The dominance of nouns and adjectives aligns with the formal and informational nature of institutional documents. These findings suggest that any automated GS must prioritize these categories to effectively evaluate translation quality. It would be interesting to go further in this part of the analysis, checking, for example, the existence and extent of nominal hypertrophies (Romana, 2009); this would require developing tools other than those used in this work. In the

[2] https://www.bbva.es/diccionario-economico.html

meantime, we can now confirm the very high degree of nominalization of these texts, which in this corpus has yielded no less than 80% of nouns and adjectives.

5.4 Lexical-Semantic Associations

The semantic clustering revealed two distinct thematic zones: regulatory compliance and public policy. This thematic segregation underscores the importance of contextual understanding in translation quality assessment. The automated GS should incorporate mechanisms to distinguish between these contexts to provide more accurate and relevant evaluations.

Dimensional reduction enhances the process, revealing thematic clusters which help to categorize texts within specialized domains, streamlining terminology and cognitive load considerations, providing clarity in translational subfields, and complementing lexical-semantic associations for comprehensive text analysis.

The results indicate that a one-size-fits-all approach is insufficient for TQA in institutional translations. Instead, a context-sensitive GS that accounts for thematic and linguistic nuances is essential. This approach will enhance the reliability of automated quality assessments and reduce the need for extensive human intervention.

6 Conclusions

Our findings support the establishment of a detailed feature set for the GS, aligned with our research objectives:

- Error Tolerance Analysis: By identifying key lexical and grammatical patterns, we can better assess the error tolerance of STs. The high frequency of specific terms and PoS categories provides a basis for determining acceptable error levels in translations.
- Quality Assessment of TTs: The detailed analysis of lexical-semantic associations and functional lexicalization offers a robust framework for evaluating the quality of TTs. By comparing these features against the GS, we can ensure that translations maintain the intended meaning and coherence.

Overall, our findings support the establishment of a conceptual reference model for evaluating institutional translations, contributing to improved QA frameworks in machine translation.

This study aimed to characterize institutional texts and develop an automated Golden Standard (GS) for evaluating translations between English and Spanish. Our analysis focused on lexical and part-of-speech (PoS) frequencies, functional and technolectal lexicalization, and lexical-semantic associations. This forms the initial phase of a broader objective to explore an artificial intelligence (AI) system employing natural language processing (NLP) technologies. The system aims to automatically assess the translation costs by measuring the person-hours required to translate English source texts. This initiative holds significance for both commercial and professional sectors, offering insights into translation production costing.

Our main results are:

1. Lexical and PoS Frequencies: We found that nouns and adjectives dominate institutional texts, comprising 80% of the parts of speech. Key terms such as "*empresa*," "*público*," and "*inversión*" were identified as frequent lexical units, highlighting their importance in the genre.
2. Functional and Technolectal Lexicalization: Common lexical combinations like "financial market policy" and "public sector" were identified, but current tools struggled to extract compound Terminological Units (TUs), suggesting a need for more sophisticated extraction methods.
3. Lexical-Semantic Associations: Semantic clustering revealed two primary thematic zones: regulatory compliance and public policy. This underscores the necessity for context-sensitive translation quality assessments.

To gauge the balance between pre-processing and human intervention, measuring the translated text error tolerance is essential (Moorkens & Guerberof 2024). Factors influencing this tolerance, as defined by Reiβ and Vermeer (1996), require meticulous examination. Our findings indicate that a detailed and context-sensitive GS is essential for accurately assessing translation quality in institutional texts. By incorporating specific lexical and grammatical patterns, as well as thematic context, we can enhance the reliability of automated quality assessments and minimize the need for extensive human intervention. The study focuses on collecting and calibrating descriptive features to characterize the reference text. Special attention is paid to the lexical density of technolectal elements in institutional texts, confirming the prevalent use of terminological units (TU), which must be distinguished from phraseological usage. While AI tools excel in detecting phraseological and TUs, extracting compound nominal units from specialized terminology remains a challenge. Further research is warranted to develop tools for efficient terminological extraction. The study also sheds light on the prevalence of nominals within PoS. Nominals dominate, comprising 80% of words, with nouns and adjectives being the most prevalent. Notably, hypertrophied nominal sequences warrant investigation, presenting avenues for future research. Parsing syntactic structures using AI tools and PoS annotation aids in extracting textual patterns. Future research could explore counting nominals in uninterrupted sequences between conjugated verbs, contributing to academic advancements. Examining the cognitive load required to read texts could enrich specialized translation practices. While this aspect aligns with disciplines like linguistic psychology and neurolinguistics, its integration into major AI language models remains speculative.

The study also delves into semantics, utilizing NLP to extract lexical-semantic fields, beneficial for textual description, translation QA, and teaching. Semantic maps derived from textual embedding and dimensionality reduction offer valuable insights, paving the way for effective textual procedures. AI's ability to embed words and texts through NLP presents unprecedented opportunities for determining textual genre and error tolerance factors. However, further quantitative studies are warranted. This research endeavors to enhance techniques and procedures in the field of translation, contributing to its overall advancement.

6.1 Limitations and Future Research

While our study provides a strong foundation for developing an automated GS, several limitations must be addressed:

- Tool Limitations: The current tools used for lexical and semantic analysis showed limitations in detecting 1-gram and compound TUs. Future research should focus on improving these tools to enhance the accuracy of the GS.
- Corpus Diversity: Our analysis was based on a specific corpus of Spanish institutional texts. Expanding the corpus to include a wider range of texts and languages will help generalize the findings and improve the robustness of the GS.

Future research should also explore the integration of more sophisticated AI techniques such as deep learning models, which nowadays are evolving at unprecedented scales (Hernández Pardo, 2023, p. 30) to further refine the GS and improve its applicability across different genres and languages. Based on our findings, we recommend several future research directions:

1. Tool enhancement for compound TU detection: Current tools showed limitations in detecting compound TUs. Future research should focus on developing more advanced methods and tools to accurately identify and evaluate these lexical units, which are critical for specialized translations.
2. Expansion of corpus diversity: To generalize the findings and improve the robustness of the GS, future studies should include a wider range of texts and languages. This expansion will help ensure that the GS is applicable across different genres and linguistic contexts.
3. Integration of advanced AI techniques: Incorporating deep learning models and other sophisticated AI techniques can refine the GS and improve its applicability. Future research should explore these technologies to enhance the accuracy and reliability of TQA frameworks.
4. Error Tolerance and quality metrics: Further investigation is needed into the specific error tolerance levels and quality metrics suitable for different types of institutional texts. This will help fine-tune the GS to provide more precise assessments.
5. Contextual and cultural nuances: Additional studies should focus on the integration of cross-cultural consistency tests and contextual understanding into the GS. This approach will ensure that translations maintain their intended meaning and coherence across different cultural contexts.

In summary, this research advances our understanding of the linguistic characteristics of institutional texts and provides a detailed framework for developing a robust and context-sensitive GS for translation quality assessment. By addressing the identified limitations and building on our findings, future research can enhance the effectiveness and reliability of TQA in professional translation contexts.

We have tried to make progress in establishing a set of features of a possible ideal frame of reference for specialized translation, especially in the institutional field. We hope this research will contribute to implementing valuable techniques and procedures for our entire field of study and work

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akbari, A., Shahnazari, M. Calibrated Parsing Items Evaluation: a Step Towards Objectifying the Translation Assessment. Language Testing in Asia, 9(1), 1-27. doi: 10.1186/S40468-019-0083-X (2019).
2. Akrami, A., Ghonsooly, B., Yazdani, M., Alami, P.M. Construction and Validation of a Computerised Open-ended Bi-functional Translation Assessment System. New Voices in Translation Studies 18 (2018).
3. Álvarez García, C. Estudio del Lenguaje de Especialidad Económico: El Lenguaje del Comercio Internacional. Entreculturas, (3), 279-290 (2011).
4. Amini, M., Ravindran, L., & Lee, K-F. Implications of Using AI in Translation Studies: Trends, Challenges, and Future Direction. Asian Journal of Research in Education and Social Sciences, 6(1), 740-754 (2024).
5. Babych, B. Automated MT Evaluation Metrics and their Limitations. Tradumàtica (12):464-470. doi: 10.5565/rev/tradumatica.70 (2014).
6. Banchs, R. E., & Li, H. AM-FM: A Semantic Framework for Translation Quality Assessment. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 153–158, Portland, Oregon, USA. Association for Computational Linguistics (2011).
7. Borja, A., García-Izquierdo, I. & Montalt, V. Research Methodology in Specialised Genres for Translation Purposes. The Interpreter and Translator Trainer 3(1), 57-77. doi:10.1080/1750399X.2009.10798781 (2014).
8. Cer, D., Manning, C. D. & Jurafsky, D. The Best Lexical Metric for Phrase-Based Statistical MT System Optimization. In Kaplan, R., Burstein, J., Harper, M. & Penn, G. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 555-563. Association for Computational Linguistics (2010).
9. Daems, J., Macken, L., Vandepitte, S. Quality as the Sum of its Parts: A Two-step Approach for the Identification of Translation Problems and Translation Quality Assessment for HT and MT+PE. In Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice, pp. 63-71 (2013).
10. Chen, N. Bridging the Unbridgeable: A Semiotic Solution to Seeking Tertium Comparationis in Linguistic Analysis. Chinese Semiotic Studies, 13(3), 255-267. doi: 10.1515/CSS-2017-0015 (2017).
11. Ebeling, S.O. & Ebeling, J., Contrastive Analysis, Tertium Comparationis and Corpora. Nordic Journal of English Studies, 19(1), 97-117. doi: https://doi.org/10.35360/njes.514 (2020).
12. Eyckmans, J., Segers, W., Anckaert, P. Translation Assessment Methodology and the Prospects of European Collaboration. Language Testing and Evaluation, 26, 171-184 (2012).
13. Faber Benítez, P. The Cognitive Shift in Terminology and Specialised Translation. Monografías de Traducción e Interpretación. doi: https://doi.org/10.6035/MonTI.2009.1.5 (2009).

14. Faber, P. & Ureña Gómez-Moreno, J. M. Specialized Language Translation. In A Cognitive Linguistics View of Terminology and Specialized Language, pp.73-92. De Gruyter Mouton (2012).
15. Fan, S. (1990). A Statistical Method for Translation Quality Assessment. Target 2(1), 43–67. ISSN 0924-1884 | E-ISSN 1569-9986 (2012).
16. Fernández Parra, M. Towards a Definition and Classification of Formulaic Language for its Translation in Specialised Texts. Collocations and Idioms 1: Papers from the First Nordic Conference on Syntactic Freezes, pp. 113-127 (2007).
17. Finch, A., Akiba, Y. & Sumita, E. How Does Automatic Machine Translation Evaluation Correlate with Human Scoring as the Number of Reference Translations Increases? In Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R. & Silva, R. (eds.) Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), pp. 2019-2022. European Language Resources Association (2004).
18. Forchini, P. & Murphy, A. C. N-grams in Comparable Specialised Corpora Perspectives on Phraseology, Translation, and Pedagogy. International Journal of Corpus Linguistics 13(3), 351-367. DOI: 10.1075/ijcl.13.3.06for (2008).
19. Gajer, M. Specialised Fully Automatic Machine Translation System Delivering High Quality of Translated Texts. TASK Quarterly: Scientific Bulletin of Academic Computer Centre in Gdansk, Vol. 13, No 4, 347-353. https://journal.mostwiedzy.pl/TASKQuarterly (2009).
20. Gil Sanromán, Í. Hernández Pardo, B, Martín Matas, P, Romana García, M. L., Innovación Docente "Trujanews": Cibercompetencias de Gestión Multilingüe. In S. Liberal Ormaechea (Coord.), J. Sierra Sánchez (Coord.), Retos y Desafíos de la Innovación Educativa en la Era Post COVID-19, pp. 679-698, McGraw Hill. ISBN: 978-84-486-3223-6 (2021).
21. Giovannotti, P. Evaluating Machine Translation Quality with Conformal Predictive Distributions. 12th Symposium on Conformal and Probabilistic Prediction with Applications, COPA 2023. https://arxiv.org/abs/2306.01549v1, https://doi.org/10.48550/arXiv.2306.01549 (2023).
22. Han, C. Translation Quality Assessment: a Critical Methodological Review, The Translator, 26(3), 257-273, doi: 10.1080/13556509.2020.1834751 (2020).
23. Han, L. Smeaton, A. & Jones, G. Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods. In Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age, pp. 15–33. Association for Computational Linguistics (2021).
24. Hatim, B. The Translation of Style: Linguistic Markedness and Textual Evaluativeness. Journal of Applied Linguistics 1(3), 229-246. doi:10.1558/japl.2004.1.3.229 (2024).
25. He, Z., Wang, X., Tu, Z., Shi, S., & Wang, R. Tencent AI Lab - Shanghai Jiao Tong University Low-Resource Translation System for the WMT22 Translation Task. In Proceedings of the Seventh Conference on Machine Translation (WMT), pp. 260–267, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics (2022).
26. Hedayati, E., & Yazdani M. Translation Quality Assessment Based on House's Model: English Translations of Iran's Supreme Leader Letters to European Youth, Psychology. doi:10.22034/EFL.2020.230069.1039 Corpus ID: 221756448 (2020).
27. Hennecke, A. Traducción y Cultura: Reflexiones sobre la Dimensión Cultural de Textos y su Importancia para la Traducción. Cuadernos de Lingüística Hispánica, 103-119. doi: 10.19053/0121053X.3681 (2015).
28. Hernández Pardo, B. La Doble Cara de la Inteligencia Artificial: Retos y Oportunidades para los Expertos en Lingüística. Puntoycoma, 180, 25-32 (2023).
29. Hernández Pardo, B. & Romana García, M. L. Traducistán 2.0: El País de la Comunicación Interdisciplinar. In Guarro, A., Area, M., Marrero, J. & Sosa, J. La Transformación Digital

en la Universidad: XI Congreso Iberoamericano de Docencia Universitaria, pp. 820-837. Universidad de La Laguna (2021).
30. Hmida, F., Morin, E., Daille, B., & Planas, E. A Bilingual KRC Concordancer for Assisted Translation Revision based on Specialized Comparable Corpora. Computer Science, 22 June 2016 (2016).
31. House, J. A Model for Translation Quality Assessment. Gunter Narr Verlag, Tübingen (1977a).
32. House, J. Translation Quality Assessment: A model Revisited. Gunter Narr Verlag, Tübingen (1997b).
33. Huertas Barros, E. and Vine, J. Assessing the Situation: Closing the Gap in Translation Assessment. The Linguist, 54(4), 22-24 (2015).
34. Jaszczolt, K. M. On Translating 'What is Said'. Tertium Comparationis in Contrastive Semantics and Pragmatics. doi logohttps://doi.org/10.1075/pbns.100.26jas (2003).
35. Kurteš, S. New Horizons for Contrastive Analysis: Grammatical Prototypes as Tertium Comparationis. School of Languages & Applied Linguistics. Centre for European & International Studies Research (2009).
36. Limón Aguirre, F. Retos al Entendimiento y la Comunicación. Tinkuy: Boletín de Investigación y Debate, 20, 92-100, ISSN-e 1913-0481 (2013).
37. Lommel, A. Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies, 109-127. doi: 10.1007/978-3-319-91241-7_6 (2018).
38. Lorente, M. Tipología Verbal y Textos Especializados. In Actas del Congreso Internacional de Lingüística: Léxico & Gramática. Lugo, 25-28 September (2000).
39. Lorenzi Zanoletty, R. Del Registro al Género: Problemas de Traducción de Expresiones Coloquiales en Textos Específicos del Sector Turístico. Quaderns de Filologia, X, 173-186. https://ojs.uv.es/index.php/qfilologia/article/view/5087 (2005).
40. Lyu, C., Xu, J. & Wang, L. New Trends in Machine Translation using Large Language Models: Case Examples with ChatGPT. https://ar5iv.labs.arxiv.org/html/2305.01181v2 (2024a).
41. Lyu, C., Du, Z., Xu, J., Duan, Y., Wu, M., Lynn, T., Fikri A., Wong, D. F., Liu, S., Wang, L. A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models. https://arxiv.org/abs/2305.01181v3 (2024b).
42. Mariana, V. R. The Multidimensional Quality Metric (MQM) Framework: A New Framework for Translation Quality Assessment, Brigham Young University – Provo 4 (2014).
43. Martínez Mateo, R., Montero Martínez, S., & Moya Guijarro, A. J. The Modular Assessment Pack: a New Approach to Translation Quality assessment at the Directorate General for Translation. Perspectives Studies in Translatology 25(1):1-31. doi:10.1080/0907676X.2016.1167923 (2016).
44. Moorkens, J. & Guerberof Arenas, A. Artificial Intelligence, Automation and the Language Industry. In Massey, G., Ehrensberger-Dow, M., & Angelone, E. In Handbook of the Language Industry: Contexts, Resources and Profiles, pp. 71-79. De Gruyter Mouton (2024).
45. Muzii, L. Quality Assessment and Economic Sustainability of Translation. Gruppo L10N, Roma (2006).
46. Nema, P. & Khapra, M. M. Towards a Better Metric for Evaluating Question Generation Systems. In Riloff, E., Chiang, D., Hockenmaier, J. & Tsujii, J. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3950-3959. Association for Computational Linguistics (2018).
47. Olohan, M. Introducing Corpora in Translation Studies. Routledge (2004).

48. Portilho, T. & Drugan, J. T. Quality in Professional Translation: Assessment and Improvement. London and New York: Bloomsbury, 2013. Revista da Anpoll. 1. 400. 10.18309/anp.v1i44.1150 (2018).
49. Portolés Lázaro, J. Marcadores del Discurso y Traducción. In García Palacios, J. (aut.) Texto, Terminología y Traducción, pp. 145-168, ISBN 8474550793 (2002).
50. Rabadán, R., Labrador, B. & Ramón, N. Corpus-based Contrastive Analysis and Translation Universals: A Tool for Translation Quality Assessment English - Spanish. International Journal of Translation, 55(4), 303-328 (2009).
51. Rei, R., Stewart, C. Farinha, A. C. & Lavie, A. COMET: A Neural Framework for MT Evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2685–2702. Association for Computational Linguistics (2020).
52. Reiβ & Vermeer. Fundamentos para una Teoría Funcional de la Traducción. Ediciones Akal (1996).
53. Rivera Trigueros, I. Machine Translation Systems and Quality Assessment: a Systematic Review. Language Resources and Evaluation, 56, 1-27. 10.1007/s10579-021-09537-5 (2022).
54. Romana García, M. L. La Sintaxis en la Traducción Económica (Inglés-Español). Universidad Pontificia Comillas. Tesis doctoral (2009).
55. Romana García, M. L. La Traducción Especializada frente a la Editorial. La Linterna del Traductor, 7, 67-75. https://lalinternadeltraductor.org/n7/traduccion-no-editorial.html (2012).
56. Sevilla Muñoz, J. Fraseología y Traducción. Revista Complutense de Estudios Franceses, 12, 431-440, ISSN 1139-9368, ISSN-e 1989-8193 (1997).
57. Subhi Khalil, G. Textual Analysis of Complex Nominals Translation Errors in Economic Texts. Journal of College of Education for Women-University of Baghdad P-ISSN: 1680-8738; E-ISSN: 2663-547X. doi: http://doi.org/10.36231/coeduw/vol31no3.13 (2020).
58. Tajvidi, Gh.R. Translation Quality Assessment. Translation Studies, 3(10), 27-40. Sid. https://sid.ir/paper/96143/en (2005).
59. Valero-Garcés, C. Contrastive ESP Rhetoric: Metatext in Spanish-English Economics Texts. English for Specific Purposes, 15(4), 279-294. ISSN 0889-4906. https://doi.org/10.1016/S0889-4906(96)00013-0 (1996).
60. Vangehuchten, L. El Uso de la Estadística en la Didáctica de las Lenguas Extranjeras con Fines Específicos: Descripción del Proceso de Selección del Léxico Típico del Discurso Económico Empresarial en Español. In 7es Journées Internationales d'Analyse Statistique des Données Textuelles, pp. 1128-1135 (2004).
61. Volk, M. & Harder, S. Evaluating MT with Translations or Translators: What is the Difference? In Maegaard, B. Proceedings of Machine Translation Summit XI: Papers, pp. 499-506 (2007).
62. Wichmann S., Holman, E. W., Bakker, D. & Brown, C. H. Evaluating Linguistic Distance Measures. Physica A, 389, 3632-3639 (2010).
63. Zarco Tejada, M. A. Predicados Complejos y Traducción Automática. Universidad de Cádiz, ISBN: 84-7786-530-2 (1998).
64. Zhang, M., Li, C., Wan, M., Zhang, X & Zhao, Q. ROUGE-SEM: Better Evaluation of Summarization Using ROUGE Combined with Semantics. Expert Systems with Applications, 237(A), 121364 (2024).
65. Zhu, L. On the Origin of the Term Tertium Comparationis. Language & History, 60, 35-52. 10.1080/17597536.2017.1293373 (2017).