RAEL: Revista Electrónica de Lingüística Aplicada
Vol./Núm.: 22/1
Enero-diciembre 2023
Páginas: 164-180
Artículo recibido: 27/07/2023
Artículo aceptado: 08/01/2024
Artículo publicado 31/01/2024
Url: https://rael.aesla.org.es/index.php/RAEL/article/view/590
DOI: https://www.doi.org/10.58859/rael.v23i1.590

Financial concepts extraction and lexical simplification in Spanish

Extracción de conceptos financieros y simplificación léxica en español

BLANCA CARBAJO CORONADO
UNIVERSIDAD AUTÓNOMA DE MADRID

ANTONIO MORENO SANDOVAL
UNIVERSIDAD AUTÓNOMA DE MADRID

This paper delves into concept extraction and lexical simplification in the financial domain in Spanish. In our approach, concept extraction involves identifying relevant terms and phrases using AI language models, while lexical simplification aims to make complex financial concepts more accessible. For this study, terms were annotated in the FinT-esp financial corpus and the mT5 neural model was used for accurate term extraction. The model yielded remarkable results: 96% of the detected terms had not been manually annotated before, showcasing its noteworthy generative capability. For lexical simplification, the paper proposes three main strategies: paraphrasing, synonym substitution, and translation, all integrated into an interactive interface that addresses the issue of sentence length. This research significantly contributes to financial concept detection and offers an effective method for simplifying financial language in Spanish.

Keywords: *specialised lexicon; financial language; automatic simplification; linguistic resource; Spanish*

El artículo examina la extracción de conceptos y la simplificación léxica en el ámbito financiero en español. La extracción de conceptos implica identificar términos y frases relevantes utilizando modelos de lenguaje de inteligencia artificial, mientras que el objetivo de la simplificación léxica es hacer que los conceptos financieros complejos sean más accesibles. Se han anotado términos en el corpus financiero FinT-esp y se ha utilizado el modelo neuronal mT5 para una extracción precisa de términos, logrando resultados notables: el 96% de los términos detectados no habían sido previamente anotados de manera manual, lo que demuestra su capacidad generativa. Para la simplificación léxica, se proponen tres estrategias principales: parafraseo, sustitución de sinónimos y traducción, integradas en una interfaz interactiva que aborda el problema de la longitud de las oraciones. Esta investigación contribuye significativamente a la detección de conceptos financieros y ofrece un método efectivo para simplificar el lenguaje financiero en español.

Palabras clave: *léxico especializado; lenguaje financiero; simplificación automática; recurso lingüístico; español*

1. Introduction: Concept extraction and simplification in a computational context

This paper addresses two closely related topics: concept extraction within a specialised domain and lexical simplification of complex concepts for a general audience. In this section we provide definitions for both.

This article is structured as follows: Section 2 presents the methodology employed in this study. Subsequently, in Section 3, we describe our approach to the neural-based conceptual extractor, including annotation guidelines and the evaluation of automatic keyword extraction results. Sections 4 and 5 focus on lexical simplification strategies and their potential issues. Our findings and concluding remarks are presented at the end.

Corpus linguistics, as defined by Parodi (2008), is a methodological approach that uses empirical study to explore language through observable data stored in electronic corpora. Parodi highlights the importance of including a diverse range of language states or varieties within a corpus, whether written or spoken, to effectively capture the essence of the language. The principles of representativeness and balance in corpus design, as detailed by Sinclair (2005), are crucial to ensure that the language patterns observed are natural and unbiased, particularly in their communicative function within specific communities, like the financial sector. Corpus linguistics, with its data-driven approach, facilitates a more objective and comprehensive analysis of language use. This analysis often employs methods like frequency analysis, concordancing, and collocation analysis, which are adept at uncovering language patterns, structures, and usage trends. One application of such methods is automatic concept extraction.

Automatic concept extraction, or keyword extraction, is a process carried out by statistical or natural language processing (NLP) tools to identify and extract the most relevant and meaningful words or phrases from a document. These keywords represent the key concepts or main themes present in the text and can aid in summarising its content or efficiently categorising it.

Prior to the advent of machine learning models, the most prevalent approach involved identifying potential terms based on their frequency of occurrence within the text. Another common approach was the recognition of specific linguistic patterns, searching for word or phrase structures that typically signify technical or specialised terminology.

One of the most popular tools for term extracting is Sketch Engine, a linguistic software tool designed for analysing and exploring language data.[1] The architecture of Sketch Engine consists of a database management system for indexing large text corpora, a web interface for corpus searches, and tools for corpus building and management.

Its renowned term extraction functionality allows users to identify and extract words and phrases typical to specific texts or corpora. It is based on statistically comparing word occurrences across two distinct corpora: a "focus corpus" containing texts from a specific field of interest, and a "reference corpus" which includes a more general range of lexicon for comparative purposes. This comparative method represents the conventional approach.

Sketch Engine distinguishes between "keywords" (individual words or tokens) and "terms" (multi-word expressions). This differentiation is based on Sketch Engine's specialised method of extracting multi-word expressions ("terms") through a term grammar. This grammar applies specific formation rules to identify potential terms in each language, such as ADJ + NOUN + NOUN (e.g., "irritable bowel syndrome") or NOUN + PREP + DET + NOUN + ADJ

[1] Accessible at https://www.sketchengine.eu/

(e.g., "síndrome del intestino irritable"). Nevertheless, we do not strictly adhere to this distinction in our research, as both terms and keywords can consist of single or multiple words. The critical aspect is that the semantic function expressed as a "typical or characteristic lexical unit" is present in both cases.

Consequently, our proposal introduces a significant change in the extraction strategy: we shift from a focus on term-formation rules and comparative frequency to using vector semantics from Artificial Intelligence language models.

Vector semantics is a technique for representing words and phrases as vectors of numerical values. There are several different methods for computing vector representations of words, but they all share a common goal: to capture the semantic relationships between words. One common approach is to use neural networks to train on a large corpus of text. The neural network learns to map words to vectors in a way that preserves their semantic relationships. Vector semantics combines two ideas: a) linguistic distributionalist intuition (as proposed by Harris and Firth in the 40s and 50s), where the meaning of a word is defined by counting what other words occur in its environment; b) vector intuition, where the meaning of a word can be expressed as a vector, a list the numbers, a point in an N-dimensional space. Mikolov, Chen, Corrado and Dean (2013) present the original model of this technique that has radically changed semantic representation in AI. Recently, Rigouts Terryn, Hoste, Drouin and Lefever (2020) or Lang, Wachowiak, Heinisch and Gromann (2021) have applied this technique to automatic concept extraction. As Rigouts Terryn, Hoste and Lefever (2022: 1) propose, "with the rise of neural networks and word embeddings, the next development in ATE might be towards sequential approaches (…), classifying each occurrence of each token within its original context".

The effectiveness of automatic concept extraction is influenced by the context and the particular goals of the analysis, as indicated by Lang et al. (2021) and Rigouts Terryn et al. (2022). Lang et al. (2021) note that the performance of various methods can be inconsistent across different domains. Furthermore, Rigouts Terryn et al. (2022) point out that the success of extraction systems depends on aspects like the frequency and length of terms, with the extraction of less common and longer terms presenting more significant challenges.

In the financial domain, automatic concept extraction is widely employed to analyse and organise large volumes of financial information, such as corporate reports, economic news, and regulatory documents. Some examples of extracted keywords in the financial context could be:

- "Stocks", relevant for stock market analysis and investment.
- "Dividends", related to dividend payments to company shareholders.
- "Inflation", referring to the general and sustained increase in prices of goods and services in an economy.
- "Exchange rate", related to the relative value of a currency compared to another. It is significant in international markets.
- "Insurance policy", associated with insurance contracts and risk coverage.
- "Investment funds", referring to collective investment vehicles, where investors pool their resources to invest in a diversified portfolio of financial assets.
- "Public debt", related to government borrowing through bond issuance and other financial instruments.
- "Profit and loss", used in financial reporting to describe a company's financial results during a specific period.

A careful analysis of these terms reveals the critical importance of lexical simplification in enhancing clarity in communication. Indeed, simplifying language is one of the key elements in ensuring effective communication within specialised languages. It involves using simpler

and more accessible language to explain complex or technical concepts, aiming to make the information understandable to a broader audience without extensive experience or knowledge in that particular domain. It can even be adapted for people with "cognitive impairments", such as people with ageing or intellectual disabilities, non-native speakers, and others with difficulties in reading and understanding information (Alarcón, Moreno & Martínez, 2023).

There are several initiatives and recommendations on simplification: *Web Content Accessibilities Guidelines (WCAG)* or *Lectura fácil: Pautas y recomendaciones para la elaboración de documentos* (UNE 153101:2018 EX) by Asociación Española de Normalización.[2] The primary recommendation is to use a simplified lexicon of complex concepts or terms.

The following are some examples of how financial concepts could be simplified:

(1) Complex concept: "Market capitalisation".
Simple language: "Size of the company in the market".
(2) Complex concept: "Credit risk".
Simple language: "Risk of someone not repaying what you lent them".

García Asensio and Montolío (2018), among many other authors, recommend "anti-baroque" (conciseness and simplicity) in vocabulary to enhance comprehension. Although financial discourse is characterised by expository brevity, i.e., the use of abbreviations, acronyms, and symbols (Román, 2016), the financial lexicon in Spanish is excessively technical, obscure and complex due to the terminological pressure of English (Mateo, 2007). It is common to find untranslated terms, lexical or structural calques and hybrid creations, significantly impeding text comprehension.

From the computational perspective, Saggion (2017) presents a comprehensive overview using ontologies and automatic sense disambiguation. In Spanish, we have two online resources, CLARA and ArText, addressing this issue for administrative texts, among other discursive genres.[3]

We aim to apply these concepts to financial language in Spanish, particularly for understanding corporate annual reports (see Gisbert, 2021).

2. METHODOLOGY: DATA AND TOOLS

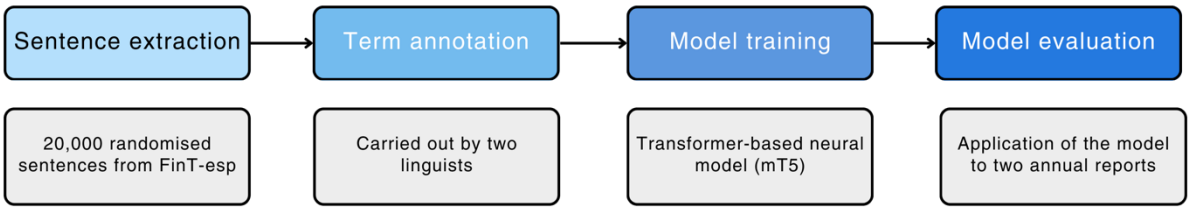

Figure 1: ***Diagram outlining the process of developing and evaluating the term recognition model***

We have used the FinT-esp corpus (Moreno-Sandoval, Gisbert & Montoro, 2020) as our data source. FinT-esp is a Spanish corpus comprising annual reports from companies listed on the IBEX.[4] The corpus comprises 388 documents, 23 million words, and 2 million sentences. Given that it encompasses a collection of updated texts (from 2014 to 2017) and is written by

[2] WCAG accessible at https://www.w3.org/WAI/standards-guidelines/wcag/
[3] CLARA accessible at https://clara.comunicacionclara.com/, and ArText accessible at http://sistema-artext.com/
[4] Information available at http://www.lllf.uam.es/es/FINT-ESP.html

and for financial specialists, it contains the jargon used by financial experts. It is suitable for terminological and conceptual extraction.

From this corpus, 20,000 sentences were randomly extracted. Four linguists manually annotated the sentences following the annotation guidelines described in Section 3.1. At least two linguists annotated each sentence, and agreement between annotators was sought after a discussion. Of the initial sentences, only 11,150 contained at least one key concept, comprising approximately 1.2 million words. The total number of keywords (types) is 5,289, distributed among 20,819 tokens, resulting in a type/token ratio (TTR) of 0.26. The variety of tokens per type is due to many multi-word concepts and inflected variants.

Using this dataset, a transformer-based neural model named mT5 (Xue, Constant, Roberts, Kale, Al-Rfou, Siddhant, Barua & Raffel, 2020) was trained. The model functions as an automatic term extractor (ATE): The model functions as an automatic term extractor (ATE): when a sentence is input, it outputs the keywords it contains. Two complete reports from the FinT-esp corpus were passed through the model to evaluate its performance, resulting in 12,564 candidate keywords (single and multi-word units). Most terms have a length of 1 to 3 tokens, though the range extends from 1 to 9 tokens.

Before reviewing the results, each term was annotated with its Part-Of-Speech (NOUN, ADJ, VERB, ADP, DET, CCONJ) using the spaCy tagger (https://spacy.io/api/tagger). Finally, the first author reviewed all terms, eliminating those that did not meet the requirements defined in the annotation guidelines (see 3.1). The result consists of 9,141 concepts, accounting for 72.9% of the initial candidates generated by the model. This accuracy rate is considerably high compared to general ATEs.

The whole process of extracting terms and concepts is depicted in Figure 1. The first two phases entail manual annotation by linguists, after which automatic extraction ensues from the lexicon compiled by specialists. The subsequent section elucidates the development of manual annotation and linguistic evaluation of automatic extraction results.

3. A CONCEPTUAL EXTRACTOR BASED ON DEEP LEARNING

In this section, we present our approach to building a conceptual extractor based on deep learning techniques. The specifics of the annotation guidelines can be found in Section 3.1, and the key findings and results of the model pertaining to this study are presented in Section 3.2.

3.1 *Annotation guidelines*

We designed an annotation guide that provides detailed instructions for annotating entities in a corpus of corporate annual reports in the financial and business domains. It is important to note that the annotation process involved random sentences from the corpus. The objective was to label relevant concepts within these reports, including linguistically specialised terms, highly specialised terminology used exclusively by experts, commonly used one-word or multi-word units, as well as recurring expressions.

For the annotation process, we used Doccano, an open-source software tool designed for annotating text in NLP projects.[5] Doccano supports collaborative work, enabling multiple users to contribute simultaneously.

Four categories were defined for the task. These include:

1) RSC (*responsabilidad social corporativa*, "corporate social responsibility"): This category includes terms related to the company's social and environmental

[5] Accessible at https://doccano.herokuapp.com/

management policies. It also includes European and Spanish directives and regulations on CSR. A few examples are *derechos de emisión* ("emission rights"), *riesgos ASG* ("ESG risks"), *responsabilidad social* ("social responsibility"), or "emission trading system".

2) GC (*gobierno corporativo*, "corporate governance"): This category includes terms related to the company's governance structures. Some examples are: *consejero independiente* ("independent director"), *consejo de administración* ("board of directors"), *pacto parasocial* ("shareholders' agreement"), or *reglamento interno de conducta* ("internal code of conduct").

This category also includes all types of committees and commissions that oversee good governance within the groups, such as *Comité de Cumplimiento Normativo* ("Compliance Committee"), or *Comisión de Retribuciones* ("Remuneration Committee").

3) EST (*estrategia y gestión*, "strategy and management"): This category includes terms related to the company's management style and strategy. A few examples include *estrategia multimarca* ("multi-brand strategy"), *integración sociolaboral* ("sociolabor integration"), *unión temporal de empresas* ("temporary joint venture"), *sociedad de garantía recíproca* ("mutual guarantee society"), *carterización* ("portfolio management"), or *estrategia de reposicionamiento* ("repositioning strategy").

This category also includes some terms directly related to employees, such as *acción formativa* ("training action") or *índice de gravedad* ("severity index"), an indicator of the severity of work accidents.

4) CONTAFIN (*contabilidad y finanzas*, "accounting and finance"): This category includes terms related to risks, financial and corporate operations, and accounting regulation and information. A few examples are *riesgo de crédito* ("credit risk"), *repreciación de activos* ("revaluation of assets"), *acuerdo de fusión* ("merger agreement"), *colocación privada acelerada* ("accelerated private placement"), or *impuesto de sociedades* ("corporate income tax").

This category also covers a series of basic accounting and finance terms frequently used in reports and annual accounts, such as *tasa* ("rate"), *impuesto* ("tax"), *inversión* ("investment"), *ingreso* ("revenue"), *provisión* ("provision"), etc. Additionally, this category encompasses stock indices like IBEX 35, Dow Jones, S&P 500, and financial authorities such as the European Central Bank (ECB) and the International Monetary Fund (IMF).

3.1.1 *Annotation rules*

1) In financial reports, it is common to encounter multiple forms of a term appearing together, such as an English term, its Spanish version, and its acronym. Each form of the term was annotated separately, excluding punctuation marks that were not part of the term. In the following example, *Fondo Único de Resolución* and *FUR* were annotated separately:

(3) (...) evitar en la medida de la posible el recurso al [Fondo Único de Resolución] ([FUR]).

"(...) avoid as much as possible the appeal to the [Single Resolution Fund] ([SRF])."

2) In cases where two terms were coordinated and one of them was truncated, both terms were annotated as a single term, including the conjunction (or comma). For instance, the string *beneficio antes de partidas excepcionales y después de impuestos* was treated as a complete term, even though it consists of two separate terms: *beneficio antes de partidas excepcionales* ("profit before exceptional items") and *beneficio después de impuestos* ("profit after tax").

However, if the two coordinated terms belonged to different categories, only the first item was annotated with its corresponding label. For example, in the case of *estructura operativa y financiera*, the term *estructura operativa* ("operational structure") was labelled as EST, while *estructura financiera* ("financial structure") was not annotated because it belongs to the CONTAFIN category.

This rule also applied when a term is coordinated with a noun that was not considered a term for this task. In those cases, only the first part was annotated. For instance, in the phrase *información financiera y no financiera*, only *información financiera*, meaning "financial information", was annotated.

This approach was adopted due to the limitations of the annotation tool used, Doccano, which does not support labelling relationships between annotated terms, such as linking the core term with each of its corresponding adjectives.

3) If the term was a verb, only the main verb was annotated in compound tenses, thereby avoiding any issues related to periphrasis. Take the following sentence as an example:

(4) (...) la Sociedad ha [amortizado] todos los activos intangibles de vida indefinida de manera retroactiva desde que dejaron de [amortizarse] tras la entrada en vigor del PGC 2007.
"(...) the Company has [amortised] all intangible assets with an indefinite life retroactively since they ceased to be [amortised] after the entry into force of PGC 2007."

4) Reduced terms due to generalisations, such as *consejo* (referring to the Board of Directors) or *comisión* (referring to any of the Governance Committees) were annotated. Similarly, terms that appeared reduced due to syntactic context were annotated accordingly.

5) Elements of a term that appeared separated due to syntactic or contextual requirements were annotated independently. For example:

(5) Los demás [arrendamientos] se clasifican como [operativos] (...).
"Other [leases] are classified as [operating] (...)."

6) Terms that are part of the names of departments within different companies were annotated with their corresponding labels. For example:

(6) Departamento de [Auditoría Interna]
"[Internal Audit] Department"

(7) Departamento de [Inversiones Financieras]
"[Financial Investments] Department"

(8) Departamento de [marketing]
"[Marketing] Department"

7) The annotation process faced a significant challenge with terms structured as N + *por* + N. Distinguishing between separate noun terms and lexicalised terms proved to be difficult. To tackle this issue, an expert in finance was consulted for guidance. It was determined that some of these terms should be annotated as a whole, while others should be annotated separately. For instance, terms like *ajuste por valoración* ("valuation adjustment"), *beneficio por acción* ("earnings per share"), or act*ivo por impuesto diferido* ("deferred tax asset") were annotated as a complete unit. On the contrary, in sequences like *[comisión] por [operaciones financieras]* ("commission for financial operations"), *[corrección de valor] por [pérdidas]* ("value correction for losses"), or *[gasto] por [impuesto devengado]* ("expense for accrued taxes") each noun and its modifiers were annotated individually.

8) Ambiguity posed another challenge in the annotation process. In certain cases, the ambiguity did not stem from potential classification into multiple categories, but rather from a term's dual function as a specialised term in specific contexts and a general word in others. Terms like *acción* ("share"), *participación* ("participation"), *compensación* ("compensation"), *aportación* ("contribution"), and *deducir* ("to deduct") were annotated exclusively when they appeared in financial or accounting contexts.

When it comes to ambiguity between categories, a common occurrence involves terms that can fall under both the CONTAFIN and EST categories. These terms can pertain to both the strategy and management field, as well as the accounting and finance domain. Examples include *gestión del riesgo* ("risk management"), *riesgo operativo* ("operational risk"), or *filial participada* ("participated subsidiary"). To address this, decisions were made to create closed lists of terms for each respective category. Moreover, the tagging of certain terms depended on the specific context. For instance, in the following sentence *cartera* should be tagged as EST:

(9) Las optimizaciones de [carteras] de clientes llevadas a cabo (...) responden al objetivo prioritario de Prosegur de mantener elevados márgenes de rentabilidad y garantizar los retornos de las inversiones.
"The optimisations of customer [portfolios] carried out (...) respond to Prosegur's priority objective of maintaining high profitability margins and guaranteeing returns on investments."

On the other hand, in this sentence, it should be annotated as CONTAFIN:

(10) En los últimos años la exposición total al riesgo soberano se ha mantenido en niveles adecuados para soportar los motivos regulatorios y estratégicos de esta [cartera].
"In recent years, total sovereign risk exposure has remained at levels adequate to support the regulatory and strategic rationale for this [portfolio]."

9) Label scope. To facilitate annotation, certain modifiers were always tagged together with the noun head in the CONTAFIN category. These modifiers hold significance as they align with accounting classifications. For example:

   - *adquirido* ("acquired"): *sociedad adquirida* ("acquired company"), *activo adquirido* ("acquired asset").
   - *arriesgado* ("in risk"): *capital arriesgado* ("risk capital", when synonymous with "venture capital")
   - *atrasado* ("overdue"): *deuda atrasada* ("overdue debt")
   - *definitivo* ("definitive"): *liquidación definitiva* ("definitive settlement")
   - *dudoso* ("doubtful"): *saldo de garantías concedidas dudosas* ("balance of doubtful guarantees"), *activo dudoso* ("doubtful asset"), *inversión dudosa* ("doubtful investment").

The same rule applies to the following structures (*a* + VERB). They are always annotated together with the noun:

   - *a cobrar* ("receivable"): *partida a cobrar* ("receivable item"), *saldo a cobrar* ("receivable balance"), *cuenta a cobrar* ("receivable account").
   - *a compensar* ("to offset"). Example: *pérdida a compensar* ("loss to offset").
   - *a cubrir* ("to cover"). Example: *pasivo a cubrir* ("liability to cover").

10) Articles, cardinal numbers, page numbers, company names, and institutions not specified in our list were not annotated.

Overall, the annotation guide provides comprehensive instructions for annotating financial and business terminology in sentences from corporate annual reports. It covers various annotation scenarios, addresses ambiguity cases, and ensures consistency in the annotation process.

3.2 *Evaluation of annotated terms with the neural model*

As mentioned in Section 2, the neural model used to identify financial terms automatically has extracted over 12,000 possible terms. After being manually reviewed by linguists, 9,141 of these terms were confirmed to be valid. These concepts are categorised into four groups: accounting and finance, strategy and management, corporate governance, and corporate social responsibility. In addition, we manually annotated 5,289 terms, which were also categorised into these four groups. Figure 2 displays the distribution of terms in each category in both sets.

However, a significant distinction exists in the methodology employed for candidate extraction. Manual detection was carried out on small excerpts (sentences) extracted from a diverse range of financial documents, while the automatically identified terms were applied to two complete financial reports.

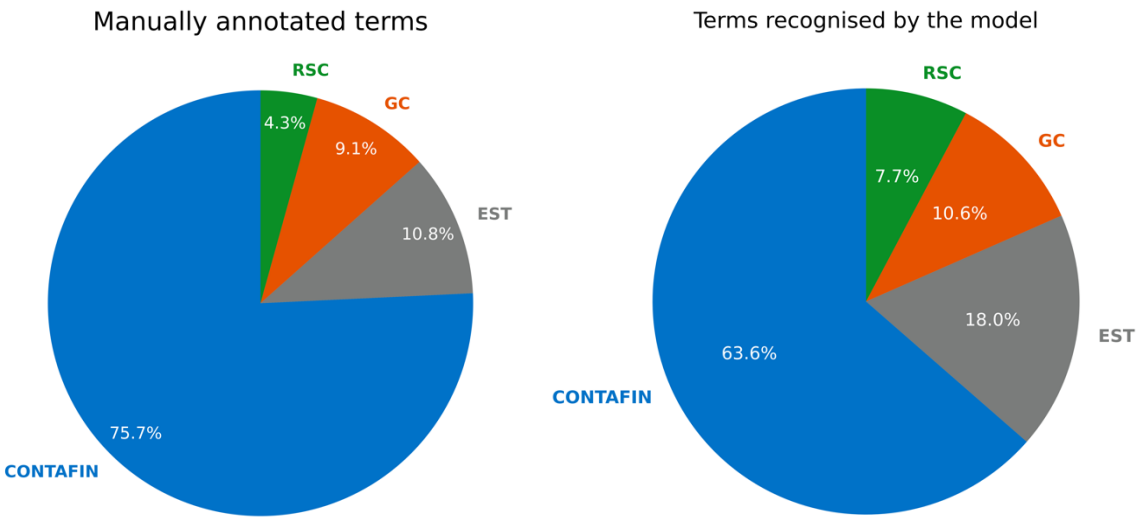

Figure 2: ***Distribution of categories in the terms recognised by the model and the annotated ones***

Upon comparing the distributions of manually annotated and automatically recognised terms, significant differences have been observed in the category distribution between the two sets. Some categories exhibit noticeable variations in frequency percentages between the two groups of terms.

The "accounting and finance" (CONTAFIN) category shows the largest difference, with 12.1% more frequency in the manually annotated terms compared to the terms recognised by the model. Conversely, the "strategy and management" (EST) category exhibits a notable discrepancy in the opposite direction, with a 7.2% reduction in the manually annotated terms compared to the terms extracted by the model. Additionally, the "corporate governance" (GC) and "corporate social responsibility" (RSC) categories also have a higher presence in the terms extracted by the model, specifically appearing with 1.5% and 3.4% more frequency, respectively, in the manually annotated terms compared to the terms recognised by the model.

Further analysis revealed that 8,776 terms extracted by the model were not found in the manually annotated list. This is illustrated in Figure 3, which shows that only 365 out of the 9,141 extracted terms were previously annotated.

The disparity in term distribution between manually extracted and automatically detected terms stems from the varying sample sources. Manual extraction involved fragments from diverse economic sectors, while automated detection focused on full reports from two banks. In other words, manually selected terms exhibit broader coverage, while automatically identified concepts are primarily focused on the specific banking domain. Consequently, the shared terms distribution might differ when considering documents from other industrial sectors.

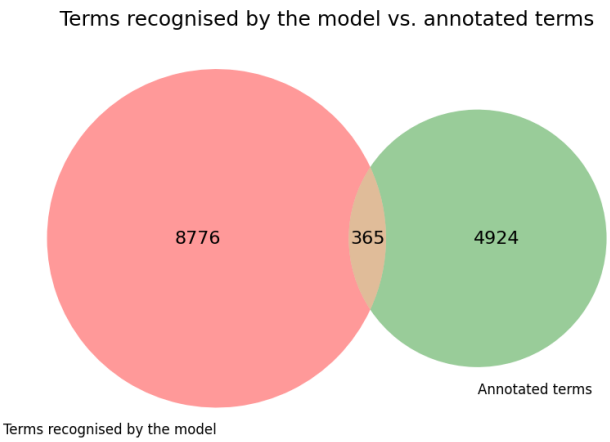

Figure 3: ***Intersection of the annotated and model-recognised term sets***

These findings are promising because they indicate that the automatic recognition model identifies new financial terms not previously included in the lexicon. This highlights the model's ability to not only recognise known terms but also discover new, relevant ones, expanding the comprehensiveness of the Spanish financial lexicon. Here are a few examples of these previously unannotated terms:

- *governance de riesgos* ("governance of risks")
- *activo generador de valor* ("value-generating asset")
- *entidad de gestión colectiva* ("collective management entity")
- *precio al riesgo* ("price at risk")
- *política de externalización de servicios* ("policy of outsourcing services")
- *valor revalorizado* ("revalued value")
- *compensación de emisiones* ("emission compensation")
- *retribución sostenible* ("sustainable remuneration")
- *accionista significativo vinculado* ("linked significant shareholder")
- *comisión de divulgación* ("disclosure commission")
- *esquema de gobierno* ("governance scheme")
- *consejero comercial* ("business advisor")
- *gestión operativa de activos* ("operational asset management")
- *sociedad absorbente* ("absorbing company")
- *dirección estratégica* ("strategic direction")
- *división de riesgos* ("risk division")
- *declaración ambiental de producto* ("environmental product declaration")
- *equipo de alta dirección* ("high-level management team")
- "general counsel"
- *sociedad holding* ("holding company")

Regarding the terms extracted by the model, it is interesting to consider the prevalent patterns among them, as these will be the ones targeted for simplification at a later stage. Figure 4 displays the distribution of the 12 most frequent POS patterns for the 9,141 financial concepts extracted by the neural model from two extensive financial reports.

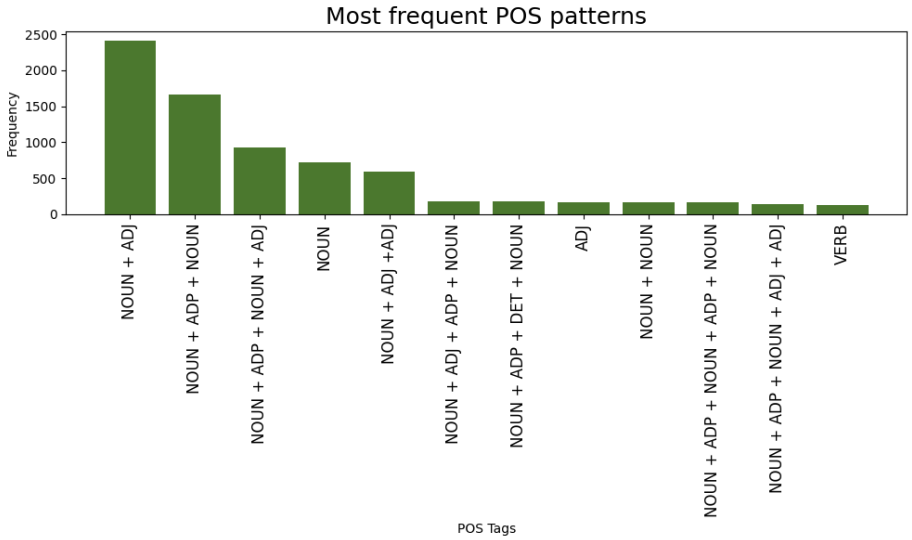

Figure 4: ***Distribution of the 12 most frequent POS patterns sorted by frequency***

This statistic is more representative than the one obtained from manual annotations carried out on 1,150 randomly selected sentences from multiple reports.

The most frequent POS pattern in financial Spanish is NOUN + ADJ with a frequency of 2,417 occurrences, followed by NOUN + ADP + NOUN with 1,661 occurrences, and NOUN + ADP + NOUN + ADJ with 927 occurrences. On the other hand, the least frequent POS pattern is VERB with only 124 occurrences.

To conclude this part of the paper, we have developed a neural Automatic Term Extractor (ATE) capable of proposing new keywords with a high accuracy rate, thus creating a conceptual database. In the subsequent sections, we will explore how a subset of this conceptual repository can be utilised for lexical simplification proposals.

4. TYPES OF LEXICAL SIMPLIFICATION STRATEGIES

To provide accurate simplifications of financial terms, they were classified into four distinct categories, based on a previous study (Vargas & Carbajo, 2021):

1) Pure Anglicisms: “backloading”, “badwill”, “cash pooling”.

2) Acronyms: *GI* (grupo de interés, “interest group”), *ANS* (*acuerdo de nivel de servicio*, “service level agreement”), *TIR* (*tasa interna de retorno*, “internal rate of return”).

3) Extended multiword terms (with four or more words): *método del exceso de beneficios multiperíodo* (“multi-period excess earnings method”), *contrato de financiación sindicada* (“syndicated loan agreement”), *unidad generadora de efectivo* (“cash-generating unit”).

4) Short multi-word terms (with less than three words): *carterizado* (“in the portfolio”, ADJ), *filialización* (“subsidiarisation”), *operación de autocartera* (“treasury stock operation”).

These categories provide valuable insights into the nature of the problem. We are faced with two separate challenges: dealing with terms originating from English and addressing terms in Spanish, each requiring a distinct approach. In the case of acronyms, providing the complete

term is insufficient; the extended terms also require simplification. The inadequacy of merely providing the full form of an acronym lies in the need for simplicity and context. Expanded acronyms often contain complex language or technical terms that may still be confusing. Providing context and simplification is crucial to ensure the information is accessible to a broader audience.

In our approach to simplifying financial terminology, we have incorporated three key strategies: paraphrasing, synonym substitution, and translation. Each strategy is carefully chosen to respect the intricacies of financial language, ensuring that the simplified terms remain accurate and contextually appropriate. This approach aligns with the insights of Cabré (2010), who highlights the critical importance of considering context and audience in lexical simplification. Furthermore, Fuertes-Olivera, Tarp and Sepstrup (2018) underscore the value of digital lexicography tools in this process.

While we provide English translations for clarity, they might not always mirror the exact register or technicality of the Spanish terms. Detailed examples of each strategy are presented to illustrate their application in real-world scenarios.

Paraphrasing and synonym substitution were employed for the terms in Spanish. Paraphrasing involves the reformulation of the original term using a different syntactic structure. Paraphrased terms tend to be longer and more syntactically complex than the original term. This technique is used when a longer explanation is needed to maintain the semantic integrity of the text. For example, the term *colateralización* ("collateralisation") can be paraphrased as *uso de activos como garantía de un préstamo*, which translates to "the use of assets as collateral for a loan". Similarly, *hipótesis actuarial* ("actuary hypothesis") can be paraphrased as *predicción que depende del análisis de los riesgos*, meaning "a prediction that relies on risk analysis". In our paraphrasing methodology, we emphasised the use of conjugated verbs over nominalisations to enhance clarity and readability. Additionally, we consciously minimised the use of excessive adjectives.

For our task, synonym substitution involves replacing the original term with another term that has an identical syntactic structure and a similar meaning This means that if the term has the structure NOUN + ADJ, the simplification should follow that same grammatical structure. For instance, the term *homogenizar* ("to homogenise") can be simplified as a *uniformar* ("to standardise"), *imputar* ("to attribute") can be substituted with *atribuir* ("to assign"), and *novar* ("to renew") can be replaced by *renegociar* ("to renegotiate"). Additionally, *inadmisión* ("inadmissibility") can be synonymous with *rechazo* ("rejection"), and *nivel de apalancamiento* ("leverage level") can be substituted with *nivel de endeudamiento* ("level of indebtedness").

For the terms in English, two strategies were employed: translation and paraphrasing. Translation involves converting English terms into their Spanish equivalents. To ensure precise and contextually relevant translations of English financial terms into Spanish, we engaged a financial specialist for expert guidance. Furthermore, our primary reference was the *Diccionario de términos económicos, financieros y comerciales* by Alcaraz Varó, Hughes and Mateo (2012).

Some examples of translated terms are "joint venture", which was rendered as *empresa conjunta*, "Chief Executive Officer" which was translated as *director ejecutivo*, and "Disclosures on Management Approach", which was translated as *información sobre el enfoque de gestión*. In some cases, cultural adaptations were made to ensure comprehension by Spanish-speaking readers unfamiliar with finance-specific terminology. For instance, "haircut" was translated as *descuento* ("discount").

Lastly, paraphrasing was employed for terms that required more explanation than a direct translation can provide. It is a reformulation of English terms into Spanish expressions that closely resemble definitions. For example, "carve-out" can be paraphrased as *retirada de*

activos que realiza la empresa cuando no está logrando sus objetivos, meaning "the withdrawal of assets made by a company when it is not achieving its objectives". Similarly, "tapering" can be paraphrased as *reducción gradual de las ayudas para impulsar la economía* ("gradual reduction of aids to boost the economy"), "back-testing" can be expressed as *técnica que analiza el éxito de la inversión utilizando datos históricos* ("a technique that analyzes investment success using historical data"), and "debit valuation adjustment" can be paraphrased as *valor actual de la deuda que no se espera pagar* ("current value of debt not expected to be paid"). Additionally, "exposure at default" can be paraphrased as *importe de deuda pendiente de pago cuando no se paga a tiempo* ("amount of outstanding debt when not paid on time"). The paraphrasing technique previously outlined for Spanish terms was similarly applied to those originally in English.

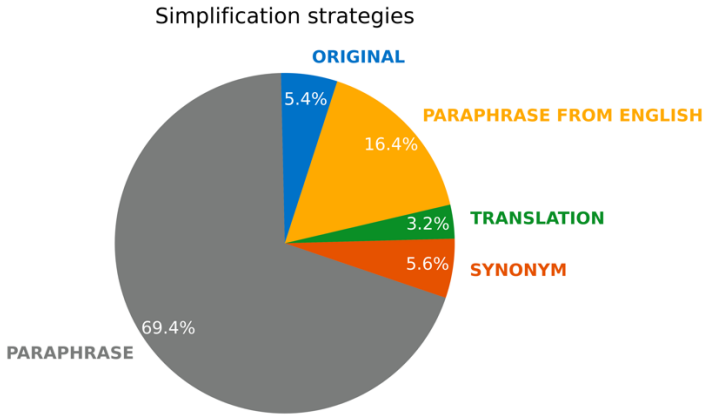

Figure 5: ***Percentage of strategies used for the 373 terms simplified***

Figure 5 shows the percentage of each strategy employed. Out of a total of 373 terms, the most prominent approach was paraphrasing, accounting for a significant majority of instances. Notably, a significant portion of terms involved paraphrasing from English. In fact, paraphrasing represented 86% of the overall strategies employed. Synonym substitution and translation were also used, albeit to a lesser extent. Interestingly, around 5% of the terms remained unchanged, primarily consisting of Spanish terms made up of three words or fewer.

The need for extensive paraphrasing instead of relying solely on synonyms when simplifying financial terms can be attributed to the specific nature of the subject matter. Financial terminology, as any other specialised domain, often requires contextual understanding for proper comprehension. Attempting to substitute terms with synonyms or hypernyms, which initially may seem like a straightforward simplification approach, runs the risk of oversimplifying and potentially distorting the intended meaning. By employing paraphrases, we provide a more comprehensive explanation that incorporates additional context, details, and examples, ensuring a clearer understanding of the concept without compromising its accuracy.

5. CHALLENGES WITH SUBSTITUTION IN THE TEXT

When it comes to simplifying financial terminology, two main obstacles were found: inflection and sentence length. While the challenge of inflection can be addressed with available solutions, the issue of sentence length poses a more persistent difficulty.

Inflection poses a challenge when the simplified term does not agree in gender or number with the original term. Additionally, the text may require that verbs are conjugated. For

example, the term *repreciación,* (“appreciation”), feminine in Spanish, becomes *ajuste del valor de un activo* (“adjustment of asset value”), masculine, losing the gender agreement, and *inadmitir* (“to reject”) becomes *rechazar* (“to reject”), which may require appropriate conjugation within the text. However, there are effective solutions available to address these challenges. Various libraries and NLP tools can automatically identify and replace terms (i.e. spaCy, NLTK). Alternatively, a pre-trained language model can be utilised, or a custom model can be trained with specific data to tackle these challenges. Moreover, personalised substitution rules can be implemented to ensure accurate and appropriate replacements.

Sentence length is the significant obstacle when generating simplified versions of financial texts. In Section 4 we showed that the strategy commonly employed to simplify financial terminology is paraphrasing, typically in the form of definitions. However, when these simplified terms are reintegrated into the original text, the resulting syntax can become more intricate and elongate the sentences. For example, the term *pacto de no competencia postcontractual* (“post-contract non-competition agreement”) may be simplified as *acuerdo de no competir con la empresa después de que el contrato termine* (“agreement not to compete with the company after the contract is completed”), resulting in a longer and more complex sentence. Similarly, *política de autocartera* (“share repurchase policy”) may be simplified as *política de una empresa sobre la compra de sus propias acciones* (“company’s policy regarding the purchase of its own shares”), further contributing to sentence length.

The production of lengthier and more complex sentences negatively impacts the readability and comprehensibility of financial texts. These longer sentences demand increased cognitive load, requiring readers to invest additional mental effort to process and understand the information. Furthermore, longer sentences can lead to information overload, as they tend to contain more information and clauses, hindering the ability of readers to extract key points effectively.

It became evident that simply breaking down the complex paraphrases we initially proposed was insufficient to address these challenges. Clearly, a shift in approach was necessary to overcome those challenges. Rather than solely focusing on translating complex texts into simpler language texts, we propose an interactive interface. Inspired by the approach taken in EASIER (Alarcón, Moreno & Martínez, 2023), the idea was to create an interface where technical terms are linked to synonyms or definitions. By clicking on an unfamiliar term within the text, the reader is presented with its synonym or definition, providing contextual support and improving both accuracy and simplicity.

6. Conclusions and future work

We have addressed two closely related topics: the automatic extraction of financial concepts and their simplification for more straightforward and accessible communication. Our approach involved manual annotation of keywords within real sentence contexts, using examples from financial annual reports.

This article presents the annotation guidelines and the distribution of keywords into four conceptual classes. From the annotated set of over 5,000 keywords, we trained a neural model based on Transformers (mT5) and evaluated its performance with two complete reports. The results were excellent, with an accuracy of 72.9%, and importantly, the model demonstrated the ability to recognise new terms not present in the initial lexicon (96%). This financial ATE in Spanish and the concept lexicon will be publicly available at the LLI-UAM webpage.[6]

For lexical simplification, we began with a subset of high-frequency keywords to tackle lexical simplification. Three different strategies were explored: paraphrasing, synonym

[6] http://www.lllf.uam.es

substitution, and translation. Paraphrasing emerged as the preferred strategy despite its associated issue of increasing syntactic complexity.

The methodology presented in this paper has yielded positive outcomes in terms of automatically extracting and simplifying complex terms. These benefits encompass enhanced information retrieval, improved comprehension, and simplified communication. By automating these tasks, the methodology streamlines workflows, reducing time and effort while maintaining consistency and accuracy. Additionally, the simplified terms facilitate understanding and engagement with complex information for individuals with diverse levels of expertise.

ACKNOWLEDGMENTS

This publication is part of the project "Computational linguistic methods for the readability and simplification of financial narratives. CLARA-FINT (PID2020-116001RB-C31)", funded by the Spanish Ministry of Science and Innovation and the State Research Agency.

The first author acknowledges the financial support provided by the FPU grant (FPU20/04007) which has been awarded by the Spanish Ministry of Science, Innovation and Universities.

REFERENCES

Alcaraz Varó, E., Hughes, B. & Mateo Martínez, J. (2012). *Diccionario de términos económicos, financieros y comerciales inglés-español, Spanish-English* (6th ed.). Barcelona: Ariel.

Alarcón, R., Moreno, L. & Martínez, P. (2023). EASIER corpus: A lexical simplification resource for people with cognitive impairments. *PLoS ONE*, *18*(4), e0283622. doi: 10.1371/journal.pone.0283622

Cabré, M. T. (2010). Terminology and translation. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of Translation Studies: Volume 1* (pp. 356-365). Amsterdam: John Benjamins Publishing Company. doi: 10.1075/hts.1.ter1

Doccano. (2024). Doccano [Software]. Retrieved from https://doccano.herokuapp.com/

Fuertes-Olivera, P. A., Tarp, S. & Sepstrup, P. (2018). New Insights in the Design and Compilation of Digital Bilingual Lexicographical Products: The Case of the Diccionarios Valladolid-UVa. *Lexikos, 28*(1). doi: 10.5788/28-1-1460

García Asensio, M. A. & Montolío, E. (2018). Cuestiones del léxico. In E. Montolío (Dir.), *Manual de escritura académica y profesional: Estrategias gramaticales y discursivas* (pp. 175-220). Barcelona: Ariel Letras.

Gisbert, A. (2021). Financial Narratives. In A. Moreno-Sandoval (Ed.), *Financial Narrative Processing in Spanish* (pp. 15-50). Valencia: Tirant lo Blanch.

Lang, C., Wachowiak, L., Heinisch, B. & Gromann, D. (2021). Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains. In C. Zong, F. Xia, W. Li & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics:*

ACL-IJCNLP 2021 (pp. 3607-3620). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.316

Mateo Martínez, J. (2007). El lenguaje de las ciencias económicas. In E. Alcaraz, J. Mateo & F. Yus (Eds.), *Las lenguas profesionales y académicas* (pp. 191-203). Barcelona: Ariel.

Mikolov, T, Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representation in Vector Space. a*rXiv*, 1301.3781. doi: 10.48550/arXiv.1301.3781

Moreno-Sandoval, A., Gisbert, A. & Montoro, H. (2020). Fint-esp: a corpus of financial reports in Spanish. In M. Fuster-Márquez, C. Gregori-Signes & J. Santaemilia Ruiz (Eds.), *Multiperspectives in Analysis and Corpus Design* (pp. 89-102). Granada: Comares.

Parodi, G. (2008). Lingüística de corpus: una introducción al ámbito. RLA. *Revista de lingüística teórica y aplicada*, *46*(1), 93-119. doi: 10.4067/S0718-48832008000100006

Rigouts Terryn, A., Hoste, V., Drouin, P. & Lefever, E. (2020). TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. In B. Daille, K. Kageura & A. Rigouts Terryn (Eds.), *Proceedings of the 6th International Workshop on Computational Terminology* (pp. 85-94). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.computerm-1.12/

Rigouts Terryn, A., Hoste, V. & Lefever, E. (2022). Tagging tems in text: A supervised sequential labelling approach to automatic term extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *28*(1), 157-189. doi: 10.1075/term.21010.rig

Román Mínguez, V. (2016). Conocimiento temático y terminológico en traducción contable (inglés-español). *Linguae Revista de la Sociedad Española de Lenguas Modernas*, *3*, 227-250.

Saggion, H. (2017). *Automatic Text Simplification*. In G. Hirst (series Ed.), *Synthesis Lectures on Human Language Technologies* (Vol. 37). San Rafael, CA: Morgan & Claypool Publishers.

Sinclair, J. (2005). Corpus and Text — Basic Principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 1-16). AHDS Literature, Language, Linguistics. Oxford: Oxbow Books.

Vargas-Sierra, C. & Carbajo-Coronado, B. (2021). Anglicisms in a Financial Corpus: exploiting resources for terminological retrieval and analysis. In A. Moreno-Sandoval (Ed.), *Financial Narrative Processing in Spanish* (pp. 99-134). Valencia: Tirant lo Blanch.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A. & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 483-498). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41