





Cross-Lingual Keyword Extraction for Pesticide Terminology in Brazilian Portuguese and English

José Victor de Souza  [Institut national de la recherche scientifique (INRS-EMT), Université du Québec, Montréal, Québec, Canada | jose-victor.de-souza@inrs.ca]


Hazem Amamou  [Institut national de la recherche scientifique (INRS-EMT), Université du Québec, Montréal, Québec, Canada | hazem.amamou@inrs.ca]


Rubing Chen  [The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong | rubing.chen@connect.polyu.hk]

Elmira Salari  [Wichita State University, Wichita, Kansas, United States | exsalari1@shockers.wichita.edu]

Reto Gubelmann  [University of St. Gallen (HSG), St. Gallen, Switzerland | reto.gubelmann@unisg.ch]


Christina Niklaus  [University of St. Gallen (HSG), St. Gallen, Switzerland | christina.niklaus@unisg.ch]


Talita Serpa  [Universidade Estadual Paulista (UNESP), São José do Rio Preto, São Paulo, Brazil | talita.serpa@unesp.br]



Marcela Marques de Freitas Lima  [Universidade Estadual Paulista (UNESP), São José do Rio Preto, São Paulo, Brazil | marcela-marques.lima@unesp.br]

Paula Tavares Pinto  [Universidade Estadual Paulista (UNESP), São José do Rio Preto, São Paulo, Brazil | paula.pinto@unesp.br]

Shruti Kshirsagar  [Wichita State University, Wichita, Kansas, United States | shruti.kshirsagar@wichita.edu]

Alan Davoust  [Université du Québec en Outaouais, Gatineau, Québec, Canada | alan.davoust@uqo.ca]

Siegfried Handschuh  [University of St. Gallen (HSG), St. Gallen, Switzerland | siegfried.handschuh@unisg.ch]

Anderson Raymundo Avila   [Institut national de la recherche scientifique (INRS-EMT), Université du Québec, Montréal, Québec, Canada | anderson.avila@inrs.ca]

 Joint Research Unit in Cybersecurity (UMR INRS-UQO), 101 Rue Saint-Jean-Bosco Gatineau (Québec) J8Y 3G5.

Received: 01 April 2025 • Accepted: 22 May 2025 • Published: 09 October 2025

Abstract Agriculture plays a crucial role in Brazil's economy. As the country intensifies its activities in the sector, the use of pesticides also increases. Hence, the risks associated with pesticide-laden food consumption have become a concern for chemistry researchers. An issue affecting regulatory standardization of pesticides in Brazil is the difficulty in translating pesticide names, particularly from English. For example, the word malathion can be translated from English to Portuguese as *malatium* or *malatião*, resulting in inconsistent labeling. This issue extends to the broader problem of translating highly technical terms between languages, in particular for low-resource languages. In this work, we investigate terminological variation in the chemistry of organophosphorus pesticides. Our goal is to study strategies for domain-specific multilingual keyword extraction. To that end, two corpora were built based on pesticide-related scientific documents in Brazilian Portuguese and English, which led to a total of 84 and 210 texts, respectively, representing the low- and high-resource languages in this study. We then assessed 6 methods for keyword extraction: Simple Maths, TF-IDF, YAKE, TextRank, MultipartiteRank, and KeyBERT. We relied on a multilingual contextual BERT embedding to retrieve corresponding pesticide names in the target language. Fine-tuning was also explored to improve the multilingual representation further. Moreover, we evaluated the use of large language models (LLMs) combined with the recent retrieval-augmented generation (RAG) framework. As a result, we found that the contextual approach, combined with fine-tuning, provided the best results, contributing to enhancing Pesticide Terminology Extraction in a multilingual scenario.

Keywords: Multilingual keyword extraction, word alignment, BERT embeddings, pesticides

1 Introduction

Agriculture is a vital component of Brazil's economy, corresponding to 5-7 % of the country's gross domestic product

(GDP), which is estimated to be around \$1.8 trillion [Lopes-Ferreira *et al.*, 2022]. A great part of the Brazilian commodity production is exported. In 2020/21, for instance, out of 137 million metric tons (mmt) of soybeans produced, 83

mmt, i.e., 75 %, were exported worldwide [Kamrud *et al.*, 2022]. The second biggest producer in the world, the United States, exported 62 mmt, i.e., 56 %, out of 112 mmt produced. Given its intensive agricultural activities, Brazil has become one of the world's largest consumers of pesticides, raising concerns about their potential harm to both the environment and human health [Lopes-Ferreira *et al.*, 2022].

This work focuses on organophosphorus pesticides, a group of substances widely used in agricultural and veterinary activities, also referred to as organophosphates [Ragnarsdottir, 2000]. Their high toxicity was first recognized in the 1930s through the synthesis of compounds later used as chemical warfare agents during World War II [Chambers and Levi, 2013; Delfino *et al.*, 2009; Kloske and Witkiewicz, 2019]. These neurotoxic agents act directly on the nervous system and are particularly harmful to mammals, including humans. By the mid-20th century, derivatives such as parathion and malathion were introduced for agricultural use, with malathion being noted for its comparatively lower toxicity [Geary, 1953]. Since then, numerous organophosphates have been synthesized and widely used in agriculture, pest control, and medicine due to their interaction with the enzyme acetylcholinesterase (AChE). In Brazil, organophosphates are commonly used in agriculture, livestock, and vector control (e.g., combating *Aedes aegypti*, the dengue mosquito). According to the National Cancer Institute (INCA), pesticide exposure can cause acute effects (e.g., irritation and nausea) or chronic conditions (e.g., respiratory issues, depression, and cancer), making their misuse a public health concern. Consequently, national research has focused on their health effects, environmental impact, and detoxification processes to mitigate their harmful impacts.

Standardizing pesticide terminology is crucial for clear communication across scientific, commercial, and regulatory domains. While pesticides have scientific names, they are also assigned common names, which are meant to be concise, unique, and universally recognized. The International Standards Organization (ISO) regulates the assignment of common names under standards 257 ISO [2018] and 1750 ISO [1981], ensuring consistency in pesticide nomenclature. However, these standards are exclusively defined in English and French, leading to challenges in multilingual standardization. This creates disparities in terminology, as speakers of other languages may lack equivalent regulated names or create inconsistent translations. In this study, we focus specifically on extracting common names of pesticides – excluding scientific names and trade names – while addressing the complexities of multilingual terminology alignment.

As noted by Pinto and Lima [2018], chemistry researchers face a lack of terminological standardization for common pesticide names in Brazilian Portuguese. Common names found in scientific papers published in Brazilian Portuguese are often variable, indicating a lack of agreement among community members on what standard to follow when translating them from English. In practice, the proliferation of different translations for the same pesticide name can lead to the misinterpretation of product labels, potentially resulting in intoxication and complicating the formulation of pesticide regulations. Notably, the morphology of pesticide names is often derived from their underlying chemical structure. Thus, the

coined name must make the molecule's main chemical group easy to identify in any language. Therefore, the variability in translating these names from English to Portuguese may hinder the accurate identification and classification of those compounds.

The term 'malathion' illustrates this problem. The name refers to a common pesticide usually translated to Brazilian Portuguese as '*malationa*', '*malatiom*', or more appropriately, '*malation*'. While the first and second are possible adaptations to the target language's morphology and orthography, they do not represent the pesticide's most important chemical group. The third one, on the other hand, indicates its correct chemical group (-thion-, -*tion*- in Brazilian Portuguese, indicating there is a phosphorus-sulfur bond in its molecular structure).

Another layer to this problem is the fact that European Portuguese historically differs from its Brazilian counterpart, which leads to translations that Brazilian researchers do not use as much, like '*malatião*', commonly used in Portugal [Souza *et al.*, 2022]. Differences between Brazilian and European Portuguese in chemical terminology have been well documented in previous studies. Finatto and Kerschner [1999] highlight that the Portuguese nomenclature system, in addition to featuring orthographic distinctions, such as '*ião*' (Portugal) versus '*ion*' (Brazil), follows a different 'nomenclature school'. This is evident in lexical choices like '*sulfureto de hidrogênio*' (Portugal) versus '*ácido sulfídrico*' (Brazil). The divergence extends to the naming of periodic table elements: while '*nitrogênio*' (N) is used in Brazil, Portugal prefers the term '*azoto*' [Finatto, 1996; Finatto and Kerschner, 1999].

In the case of pesticides, as demonstrated by Souza [2023], variation in Brazilian Portuguese is also found in spelling variants (e.g., '*diclorvos*', '*diclórvos*', and '*diclorvós*'), in morphosyntactic variants (e.g., '*metil paration*', '*paration metílico*', and '*metilparation*'), and in the coexistence of domesticated terms with English loans (e.g., '*clorpirifós*' and '*chlorpyrifos*'). This dynamic and continuously evolving terminology presents a significant challenge for both linguists and computational researchers interested in retrieving and describing specialized language. Consequently, developing methods to facilitate the retrieval and translation of terms across languages boosts efforts to investigate and standardize terminology in Brazilian Portuguese.

With this in mind, this work addresses multilingual keyword extraction, defined as the task of identifying single- or multi-word expressions that capture the main topics of a document [Carrión and Casacuberta, 2022] in different languages. Given the vast number of documents published online, indexing and retrieving them remains a significant challenge [Campos *et al.*, 2020]. Keyword extraction can support various natural language processing (NLP) tasks, including information retrieval and recommendation systems [Firoozeh *et al.*, 2020]. In this study, we address keyword extraction in a domain-specific context, aiming to identify pesticide names in scientific documents written in Brazilian Portuguese. Our broader objective is to reduce mistranslations of pesticide names. To this end, we investigate methods for retrieving Brazilian pesticide names from such documents by leveraging seed keywords in English.

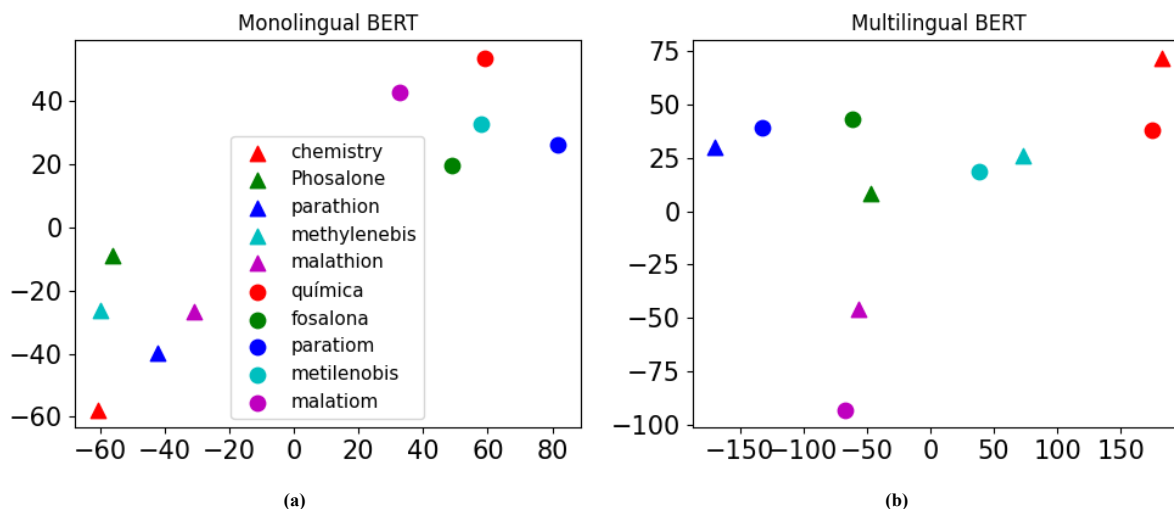


Figure 1. t-SNE visualization of the embedding space of BERT for Portuguese and English word pairs. Each point is a different instance of a given word. This plot suggests that multilingual BERT already provides some level of alignment out-of-the-box for pesticide names.

Although many efforts have been made toward domain-specific keyword extraction [Abulaish *et al.*, 2022], multilingual scenarios have not been explored enough yet, especially for Brazilian Portuguese. Existing studies focus on domain-specific texts [Sammet and Krestel, 2023; Hulth *et al.*, 2001] while overlooking multilingual scenarios. On the other hand, studies on multilingual keyword extraction are often based on general-purpose datasets [Piskorski *et al.*, 2021] and are not meant for specific domains, such as those involving pesticide names. Thus, we aim to fill this gap by tackling domain-specific multilingual keyword extraction.

Our analysis encompasses Brazilian Portuguese and English, covering 84 texts in the former and 210 in the latter. By employing English as the source language, six keyword extraction methods (TF-IDF, YAKE, TextRank, Multipartite, Simple Maths, and KeyBERT) were evaluated and compared against a ground truth keyword list. To identify corresponding pesticide names in the target language, we use multilingual contextual embeddings. Figure 1 provides a visualization of the embedding space of pesticide names, where triangles represent instances of English words and circles represent the corresponding Portuguese word occurrences. We can observe that embeddings obtained from a multilingual BERT model offer some level of alignment between Portuguese and English word pairs, a visible improvement over representations generated from monolingual BERT models. Nevertheless, to further improve word alignment, the multilingual BERT representation was fine-tuned by using synthetic technical texts specific to the pesticides domain. These embeddings were explored on two systems for multilingual keyword retrieval, which are detailed in Section 4.

Additionally, we explore the potential of large language models (LLMs) combined with retrieval-augmented generation (RAG) [Lewis *et al.*, 2020]. We hypothesize that the external knowledge introduced through RAG enhances LLMs’ capacity to answer domain-specific questions, even when the required information is absent from the knowledge encoded during training. Moreover, recent studies have shown that RAG can be more efficient than fine-tuning as a strategy to handle new knowledge or domain-specific scenarios

[Soudani *et al.*, 2024]. Therefore, two approaches are investigated, naive RAG and GraphRAG [Edge *et al.*, 2024], as options for multilingual lexical extraction in the domain of pesticides.

Thus, the main contributions of this paper are the following:

- **Organophosphorus pesticide multilingual corpora:** Development of two corpora, based on Brazilian Portuguese and English, specialized in Organophosphorus Chemistry. These corpora are publicly available for the research community interested in assessing domain-specific multilingual keyword retrieval;
- **Comparative analysis of keyword extraction methods:** Six methods for domain-specific keyword extraction were compared. This comparative analysis can help researchers select methods for retrieving pesticide names, which is relevant in scenarios where experts are available;
- **Thesauri for organophosphorus:** List of Brazilian Portuguese and English terms related to pesticide names and retrieved from the developed corpora. This list can be used as a benchmark by researchers willing to evaluate their methods for extracting organophosphorus chemistry keywords;
- **Contextual multilingual keyword extraction method:** Three methods for extracting keywords from a target corpus are presented: (1) a basic method, (2) a contextual retriever, and (2) a method based on LLM and external knowledge. These methods enable the extraction of target keywords via seed keywords from the source corpus.

The remainder of this paper is structured as follows. Section 2 presents the main related work in the field. Section 3 details the corpora development. Section 4 presents materials and methods, and Section 5 gives details on our experimental setup. Section 6 provides the results along with a linguistic evaluation. Section 7 addresses limitations related to training biases. Finally, Section 8 concludes the paper and offers insights for future research.

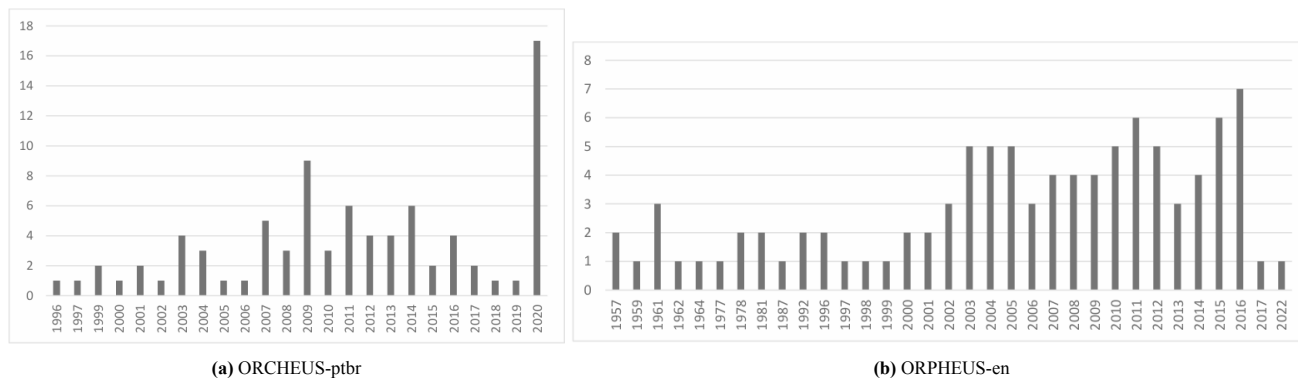


Figure 2. Number of documents by year

2 Background and Related work

2.1 Multilingual Word Alignment

Multilingual word alignment is the task of aligning word embeddings of two or more languages. This is typically achieved by making the distributions of multiple languages as close as possible in the semantic space [Och *et al.*, 1999]. A successful alignment enables word translations between two languages, given that similar representations are achieved for words conveying similar meaning (e.g., chemistry and *química*, as shown in Figure 1). This is crucial for building multilingual NLP systems [Al-Rfou’ *et al.*, 2013], such as machine translation and other cross-lingual tasks [Hämmerl *et al.*, 2024]. In general, training takes source and target languages with respective source and target embeddings to approximate the distributions of embeddings of similar words while making the distributions of dissimilar ones more distant [Xia *et al.*, 2015]. There is extensive literature on multilingual word alignment. For instance, Al-Rfou’ *et al.* [2013] trained word embeddings for more than 100 languages using their corresponding Wikipedia instances. The training aims to adjust the embeddings to improve the task of distinguishing corrupted phrases from original ones. More recently, Schuster *et al.* [2019b] presented a method to create a multilingual space of contextual embeddings by using context-independent embedding anchors to guide alignment. This approach simplifies the mapping process, compresses the space, and supports the use of bilingual dictionaries for supervision. After aligning the anchors, the mapping can be extended to the full contextual embedding spaces. Cao *et al.* [2020] used English as a target corpus (or reference embedding space) to adapt source languages. The mapping is attained using parallel data and a rotation mechanism that enables embeddings from other corpora to match the English embedding space. Individual matrices, W , must be learned for each language and applied to each word vector representation to minimize the distance between parallel word pairs.

2.2 Multilingual Keyword Extraction

Multilingual keyword extraction, the task of identifying and extracting key phrases across languages, is a crucial process in information retrieval and natural language processing. However, morphological, lexical, and syntactic differences between languages make this task more challenging

than monolingual keyword extraction [Verma *et al.*, 2022], requiring advanced cross-lingual information techniques. In general, the methods for keyword extraction are divided into two categories: supervised and unsupervised. Supervised methods, such as KEA [Witten *et al.*, 1999], rely on labeled training data, while unsupervised techniques typically rely on frequency statistics. The latter can be categorized into statistic-based models like TF-IDF and simple maths [Kilgarriff, 2009]; graph-based models like TextRank [Mihalcea and Tarau, 2004] and SingleRank [Wan and Xiao, 2008]; and deep learning-based models like CopyRNN [Meng *et al.*, 2017] and KeyBERT [Grootendorst, 2020]. Recent developments have enabled the cross-lingual semantic understanding of text, improving the accuracy of multilingual keyword extraction. For example, new datasets like MAKED [Verma *et al.*, 2022] have been created to enhance the performance of supervised methods in multilingual settings. Additionally, the advancements in unsupervised methods include the development of transformer-based models such as multilingual BERT [Pires *et al.*, 2019] and the introduction of novel algorithms based on TF-IDF [Hashemzadeh and Abdolrazzagheh-Nezhad, 2020].

2.3 Domain-specific Word Alignment

In domain-specific word alignment, researchers have explored techniques to enhance alignment quality by incorporating external knowledge sources. Domain-adapted word embeddings based on fine-tuning transformer-based models with specialized corpora have shown promising results in fields such as biomedical NLP and legal text processing [Zhu *et al.*, 2021; Schuster *et al.*, 2019a]. In particular, domain-specific BERT variants, such as BioBERT [Lee *et al.*, 2020] for biomedical text, SciBERT [Beltagy *et al.*, 2019] for scientific literature, and LEGAL-BERT [Chalkidis *et al.*, 2020] for the legal domain, have demonstrated superior performance in capturing domain-specific terminology compared to general-purpose models. Its Brazilian Portuguese counterpart, LegalBert-pt [Silveira *et al.*, 2023], similarly outperforms general language models in tasks such as named entity recognition and text classification. For low-resource languages, hybrid approaches combining statistical word alignment with transformer-based embeddings have also been proposed [Ruder *et al.*, 2021]. These methods leverage bilingual lexicons and dictionary-based constraints to refine alignment

results, addressing inconsistencies that arise in specialized domains. In our work, we build on these advances by integrating multilingual BERT embeddings with a keyword extraction pipeline to align pesticide names across languages.

Despite all the advances toward domain-specific keyword extraction [Abulaish *et al.*, 2022], multilingual scenarios have not yet been explored, especially for Brazilian Portuguese. This work aims to fill this gap. We contribute to a growing body of research on domain-specific multilingual NLP by focusing on terminology extraction for low-resource languages.

3 Organophosphorus Corpora

This section describes the two organophosphorus pesticides corpora used in this work. We start by presenting the Portuguese corpus, followed by the English one. Both corpora were compiled by Souza [2023].¹

3.1 Organophosphorus Corpus in Portuguese

The Organophosphorus Chemistry Corpus of Academic Brazilian Portuguese (ORCHEUS-ptbr) was manually collected between 2016 and 2022 from research databases, such as Athena (Unesp), Google Scholar, SciELO, and the Brazilian Digital Library of Theses and Dissertations (BDTD). The keywords used in our searches include ‘*agrotóxico*’, ‘*pesticidas*’, ‘*organofosforados*’, and ‘*organofosfato*’ (the last one being a less frequent synonym for ‘*organofosforado*’). The term ‘*agrotóxico*’ was proposed by Brazilian expert Adilson Paschoal [Moncau, 2023] to highlight both the primary human activity in which these compounds are used (agriculture) and their inherent toxicity. In English, ‘*pesticide*’ is the preferred term. As a result, the corpus comprises 84 academic texts published in scientific journals and repositories of theses/dissertations between 1996 and 2020. It holds 830,144 word occurrences and 621,213 word types (i.e., unique words), as described in Table 1. Figure 2a presents the distribution of these documents by year of publication. As the chemical nomenclature system in European Portuguese has historically differed from that used in Brazilian Portuguese, our study focuses exclusively on texts written in Brazilian Portuguese. Since the variation in terminology was initially observed in scientific articles, we prioritized academic discourse, though future work may explore less specialized registers.

3.2 Organophosphorus Corpus in English

Variation in English, albeit present, is more regulated. This is primarily because IUPAC (International Union of Pure and Applied Chemistry) and ISO (International Organization for Standardization) publish their nomenclature recommendations in English, which holds a prestigious status within the international scientific community. Consequently, in this study, we do not distinguish between English varieties, treating the language as a *lingua franca*.

Table 1. Statistics on two organophosphorus pesticides corpora

Text type	ORCHEUS-ptbr		ORPHEUS-en	
	docs	tokens	docs	tokens
Scientific paper	65	303,166	202	2,901,000
Master’s thesis	10	214,441	5	24,383
Doctoral dissertation	6	149,297	1	197,909
Undergraduate thesis	2	46,173	0	0
Report	0	0	1	95,848
Book	1	117,067	1	252,860
Total	84	830,144	210	3,472,000

Accordingly, we introduce our second corpus, in English, the Organophosphorus and Phosphorus Chemistry Corpus of Academic English (ORPHEUS-en). The texts were retrieved from scientific databases, such as Google Scholar, Athena, and ACS Publications, and the keywords used in this search were ‘*phosphorus*’, ‘*organophosphorus*’, ‘*organophosphate*’, and ‘*pesticide*’. The corpus comprises 210 texts published between 1957 and 2022, with a total of 3,472,000 tokens and 2,221,494 unique word types, as described in Figure 2b. Statistics related to this corpus are also presented in Table 1.

The English corpus is larger than the Portuguese one for two main reasons. First, academic production in STEM (Science, technology, engineering, and mathematics) fields is predominantly in English due to its status as the global *lingua franca* of science and technology. Most high-impact journals, conferences, and research institutions prioritize English for publication, fostering broader dissemination and international collaboration. As a result, there is a significantly larger volume of scholarly work available in English compared to Portuguese. Second, we also included texts on phosphorus and its inorganic derivatives in our English corpus to account for variations in substance names and phosphorus functions. This approach increases the likelihood of identifying texts on compound synthesis, thereby expanding the corpus’s nomenclature range. However, the same was not feasible for ORCHEUS-ptbr, given the relatively limited scientific output on phosphorus chemistry in Brazil compared to the international landscape. Nonetheless, the difference in corpus size reflects the challenging reality of developing methods for low-resource languages, which is the main motivation for this study.

Note that the difference in time windows between English and Brazilian Portuguese corpora reflects the earlier and broader development of pesticide chemistry research in English. In contrast, Portuguese publications emerged later alongside the field’s growth in Brazil. However, this temporal gap does not affect our study, as our methods focus on aligning domain-specific terminology across languages, independent of publication dates, provided the corpora accurately represent domain language use.

3.3 Language Pairs For Fine-tuning

The two corpora described in the previous section contain no sentence pairs. Thus, to fine-tune the multilingual BERT, we created a synthetic parallel dataset with corresponding sentences in Brazilian Portuguese and English. For that, we sourced the organophosphorus corpora of research papers within the pesticides domain, as described in the previous subsections. This yielded a total of 97,132 English sen-

¹ Available at <https://repositorio.unesp.br/entities/publication/715bd184-1c3f-441d-a514-8f0d8ffce15a>.

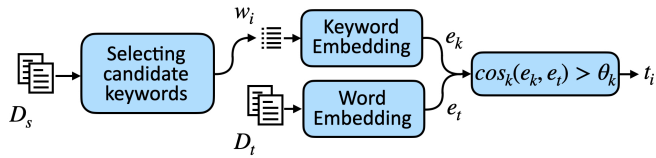


Figure 3. Basic approach for multilingual keyword retrieval for pesticide names

tences and 37,731 Portuguese sentences after pre-processing. The pre-processing involved segmenting the texts into individual sentences to facilitate effective alignment. To generate the parallel sentence pairs used for training, we employed the LLaMA-3.1-8B-Instruct language model to translate the sentences bidirectionally: English sentences were translated into Portuguese, and Portuguese sentences were translated into English. This approach allowed us to create a substantial bilingual dataset by effectively doubling the corpus through high-quality machine translation. The resulting dataset serves as a robust foundation for fine-tuning, enhancing its capability in multilingual word alignment tasks between English and Portuguese.

4 Material and Methods

In this section, we present three approaches for the domain-specific multilingual keyword extraction task explored in this work. We start with a general overview of the problem, followed by a description of the methods investigated.

4.1 General Problem Definition

Given a set of words $S = \{s_1, s_2, \dots, s_n\}$ from a source corpus of a high-resource language with vocabulary, V_s , our primary goal is to extract a list of keywords, w_i , and key-phrases, p_j , with $i = 1, 2, \dots, I$ and $j = 1, 2, \dots, J$, for a set of documents $D_s = \{d_1, d_2, \dots, d_L\}$, present in the source corpus. The keywords and key-phrases are expected to represent the main idea behind a text [Sharma and Li, 2019]. For a multilingual scenario, given a seed keyword in the source corpus, a corresponding list of keywords, t_i , must be retrieved from the documents, $D_t = \{t_1, t_2, \dots, t_f\}$, in a target corpus of a low-resource language with vocabulary, V_t . This challenge is amplified in domain-specific keyword extraction between high- and low-resource languages, where equivalent keywords may be absent from the target or source corpora. Moreover, it is often possible to find multiple equivalent keywords in the target corpus, and vice versa.

4.2 Proposed Approaches for Multilingual Keyword Extraction

We introduce four approaches for extracting terms in Brazilian Portuguese given a seed keyword in English.

4.2.1 Basic Multilingual Keyword Extraction

Figure 3 depicts a basic pipeline for the task as described in section 4.1. First, candidate keywords are retrieved from the source documents, D_s . Then, a list of domain-specific keywords, s_k , is selected based on expert assessment. The

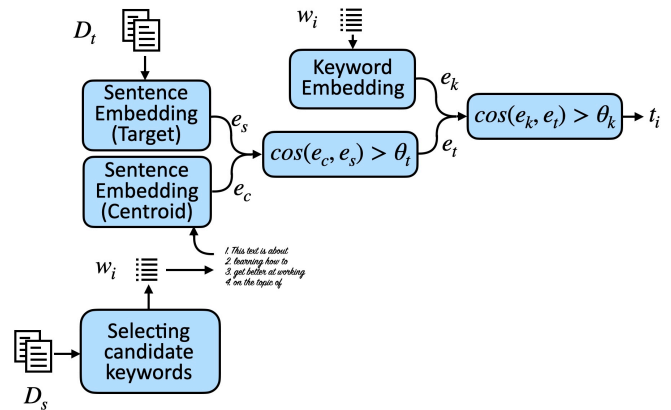


Figure 4. Proposed approach for multilingual keyword retrieval for pesticide names

selected keyword seeds are embedded using a multilingual BERT-like representation. These embeddings, referred to as e_k , are used to retrieve keywords from the target documents, e_t . For that, a cosine similarity is computed between e_k and e_t . Note that a threshold, θ_k , is used to control the level of similarity between source and target keywords. Lower thresholds lead to longer lists, including general words, while higher thresholds lead to shorter lists with more specific words.

4.2.2 Contextual Multilingual Keyword Extraction

A contextual approach is also explored as depicted in Figure 4. For that, after a set of keywords is selected from candidate ones, a list of key-phrases, p_j , is retrieved. These sentences represent the context in which each keyword appears in the source documents, D_s . We then embed these sentences and compute their centroid, as below:

$$e_c = \frac{1}{j} (e_1 + e_2 + \dots + e_j) \quad (1)$$

Note that the centroid can be seen as the multidimensional version of the mean, extracted from a set of vectors. We specifically choose the centroid representation because it provides the most central point in the vector space that minimizes the distance to all context vectors, making it more robust to outliers than a simple average. Then, this single vector representation is obtained by ensuring that it has the minimum sum of squared distances to each of the vectors in the set. The next step is to retrieve similar phrases from the target documents, D_t , by computing the cosine similarities between this centroid and embedded target sentences, e_s , representing the embedding of the s -th sentence from the target documents. This step leads to sentences from the target corpus that are relevant to the context found in the source documents and represented by e_c . As in the previous approach, the similarity between source and target sentences is controlled by a threshold, namely θ_t . The final step is to retrieve the words from the selected target sentences that are similar to the keywords from the source documents. For that, a cosine similarity is computed between the keyword, e_k , and the word embedding, e_t , as shown in Figure 4.

RAG prompt for target keyword extraction

You are a target keyword extractor for the domain of chemistry, specialized in pesticide names. Your task is to reliably retrieve **one and only one keyword** from the target corpus based on keywords from the source corpus.

Follow these rules:

1. Extract the most relevant keyword from the target context that matches the input keyword.
2. If no exact match is found, return the closest possible keyword that relates to the input.
3. Do not include any additional terms, explanations, or repetitions.
4. If no relevant keyword is found, return "N/A".

Example:

SEED KEYWORD: phosphorus

SOURCE CONTEXT: phosphates during combustion. Considering the abundance of phosphorus.

TARGET CONTEXT: Fosfatos durante a combustão. Considerando a abundância de fósforo.

OUTPUT: fósforo

Template:

SEED KEYWORD: {input_word}

SOURCE CONTEXT: {Passage Text}

TARGET CONTEXT: {Passage Text}

OUTPUT: {output_text}

Figure 5. Prompt used to give instructions to the LLM for target keyword retrieval

4.2.3 Fine-tuning for Multilingual Keyword Extraction

The key objective of multilingual word alignment is to align the semantic space distribution between languages. This is achieved so that similar embeddings from different languages should convey similar meanings. As mentioned before, this work focuses on the chemistry domain, specifically the standardization of pesticide names. Such terminology may not be seen during the pre-training phase of multilingual LLMs, such as BERT. Thus, fine-tuning these models for pesticides is expected to improve performance. For that, the two corpora described in Section 3 are used. Additionally, the synthetic data described in Section 3.3 was used to obtain a balanced dataset of sentence pairs between source and target languages.

4.2.4 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a framework that improves LLMs' generative capabilities through the use of external knowledge. By giving LLMs specific instructions, it can be seen as a type of fine-tuning without the requirement of parameter optimization via backpropagation [Lewis

et al., 2020]. It was first introduced to reduce hallucinations and enable LLMs to have access to new knowledge not seen in training, which may be the case for pesticides. Thus, we aim to find target keywords by providing the LLM with specific instructions based on source and target documents. By doing so, the LLM is exposed to domain-specific knowledge about pesticides. Figure 5 provides details concerning the template used for this task. Note that a complete example, including a seed keyword, source and target contexts, and the corresponding response with the target keyword, is provided. Following this, the template is presented to the LLM, which incorporates the source keyword, source and target contexts, and a designated response field to be completed by the LLM. We explored both a standard RAG solution and GraphRAG, the latter combining a knowledge graph with retrieval-augmented generation to facilitate information extraction from long documents, such as research articles and books. The reader is referred to the work presented in Edge et al. [2024] for detailed information.

5 Experimental Setup

We start this section by presenting the pre-processing steps. We then describe the six methods used for word extraction. Subsequently, we present the methodology to create the lists of pesticide names used as ground truth. We finalize by describing the evaluation metrics used to assess the proposed solutions.

5.1 Pre-processing

The pre-processing pipeline consisted of removing metadata such as tags, URLs, emails, and lines with affiliations or publisher names. Additionally, common abbreviations are masked to prevent incorrect tokenization. The resulting text is then tokenized using the whitespace-based method by splitting the text on spaces and then removing the surrounding punctuation. This pre-processing step standardizes the target text documents for the basic approach of keyword retrieval.

5.2 Keyword extraction methods

We examine statistical, graph-based, and BERT-based methods for automatic keyword extraction from texts.

Simple Maths: a corpus-based approach to keyword extraction by using a reference corpus [Kilgarriff, 2009]. This method allows the analysis to focus on either more common or rarer keywords. It assigns a 'keyness' score to each item in the study corpus by comparing its normalized frequency to that in a reference corpus. In this study, we used general-language corpora in Portuguese and English as reference corpora. Specifically, we employed ptTenTen18², containing 8,731,838,327 tokens, and enTenTen20³, comprising 43,125,207,462 tokens. Both corpora consist of texts collected from the web.

²<https://www.sketchengine.eu/pttten-portuguese-corpus/>

³<https://www.sketchengine.eu/ententen-english-corpus/>

	Pesticides	Non-Pesticides	
Positive	TP	FP	$precision = \frac{tp}{tp + fp}$
Negative	FN	TN	
	$recall = \frac{tp}{tp + fn}$		

Figure 6. Confusion matrix for keyword extraction evaluation.

TF-IDF: it is the product of the term frequency and the inverse of the number of documents the term appears. The first term captures the count of certain words by document. The second term in the product aims to give more weight to words that appear in fewer documents. By doing so, the measure minimizes the relevance of words with high frequencies in several documents and that, therefore, carry less discriminative information. TF-IDF may require a large corpus to perform well [Campos *et al.*, 2020]. Hence, other approaches can be found in the literature, such as the one we shall present next.

YAKE: this algorithm relies on five steps to extract keywords. The first one focuses on text pre-processing, resulting in a list of sentences. The sentences are attained using rule-based segmentation. This step also provides candidate terms. In the second step, statistical features are defined based on the previous output. These features are used to compute the score in the third step. Small term scores represent more significant 1-grams. The fourth step comprises sliding windows of arbitrary size n , which are used to generate 1-gram to n -gram terms. Candidate keywords are attained and ranked in the fifth step. The reader can find more details on this method in Campos *et al.* [2020].

TextRank: the model relies on ranking algorithms to determine the importance of a vertex within a graph [Mihalcea and Tarau, 2004]. The vertex can represent a word or a sentence. The importance of a vertex is determined by considering global information recursively extracted from the entire graph using a ‘voting’ or ‘recommendation’ system. If a vertex links to another, a vote is cast for the connected vertex, which increases its relevance. The higher the number of votes a vertex receives, the greater its importance in the graph. Thus, the vertex with the highest scores represents the most relevant keywords or key phrases within a given text.

Multipartite: in this method, a graph representation of the document is first built. As in the previous method, a ranking algorithm is used to assign a relevance score to each key phrase, which is represented by nodes in the graph [Bougouin *et al.*, 2013]. Key phrase candidates are represented in a single graph to exploit their mutually reinforcing relationship to improve candidate ranking.

KeyBERT: KeyBERT is a keyword extraction framework that utilizes BERT-based embeddings to identify key terms and phrases in a given text. It computes document embeddings using transformer-based language models and ranks candidate keywords based on their cosine similarity to the overall document representation. The tool allows for cus-

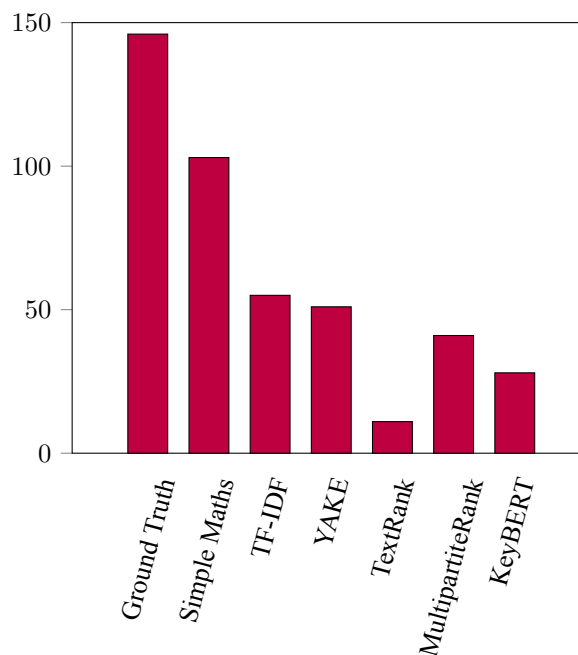


Figure 7. Number of correct keywords retrieved by each keyword extractor for the ORPHEUS-en corpus.

tomization through n -gram selection, part-of-speech filtering, and integration with external keyword extraction methods. KeyBERT is commonly applied in tasks such as topic modeling, information retrieval, and text summarization, providing a flexible approach to extracting relevant terms from unstructured text [Grootendorst, 2020].

5.3 Ground Truth

To evaluate the outcome of our experiments, we use two lists of keywords. One represents pesticide common names found in the source corpus, i.e., the ORPHEUS-en, and one is extracted from the target corpus, i.e., ORCHEUS-ptbr. All ground truth keywords were identified through a rigorous, iterative process that combined expert-driven term identification via document reviews with corpus linguistics techniques. Keywords were initially extracted from the English and Brazilian Portuguese corpora using the best-performing method identified (see Section 6). These keywords were then reviewed by a cross-disciplinary team consisting of a pesticide chemistry specialist, a translator, a terminologist, and a corpus linguist, who selected candidate organophosphorus pesticide terms. To analyze the terms in context and identify any relevant terms that may not have appeared in the keyword list, concordancers (text visualization tools that display a keyword in its immediate context) were employed. Whenever a new term was identified during this process, additional concordance lines were examined to ensure accurate interpretation and validation. This iterative process allowed for a thorough term extraction by combining multiple tools. External databases supported cases lacking sufficient in-corpus information. The terms found were validated through databases such as the Compendium of Pesticide Common Names⁴ by the British Crop Production Council (BCPC), Common Chemistry⁵ by the Chemical Abstracts

⁴<http://www.bcpcpesticidecompendium.org/>

⁵<https://commonchemistry.cas.org/>

Table 2. Comparison of keyword extractors’ performance in English and Portuguese

Method	English			Portuguese		
	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
Simple Maths	0.99	0.70	0.82	0.87	0.65	0.74
TF-IDF	0.89	0.17	0.28	0.82	0.31	0.44
YAKE	0.02	0.01	0.01	0.01	0.18	0.01
TextRank	0.16	0.03	0.05	0.05	0.12	0.07
MultipartiteRank	0.01	0.01	0.01	0.03	0.09	0.04
KeyBERT	0.89	0.06	0.11	0.62	0.11	0.18

Service (CAS), and PubChem⁶ by the National Institutes of Health (NIH). This integration allowed for a more efficient and scalable extraction of relevant terms while ensuring domain-specific accuracy through expert validation. The English ground truth comprises a list of 146 pesticide common names, while the Portuguese list consists of 200 names, both considering naming variants for the same pesticide. As discussed in Section 1, linguistic variation is prevalent in low-resource languages where standardization is still lacking, such as Brazilian Portuguese, which explains the higher number of common names found in a relatively smaller corpus.

5.4 Evaluation Metrics

As described in 5.3, a list of pesticide names is used as ground truth. This list is compared to a list of predicted keywords. Figure 6 shows the confusion matrix used to evaluate our systems, where *tp* stands for true positive and is the number of predicted pesticide names (*positive*) that can be found in the ground truth list (*pesticides*), whereas *fp* represents false positives and is the number of predicted pesticide names that cannot be found in the ground truth list (*non-pesticides*). False negatives are represented by *fn* and are the number of pesticide names in the ground truth list (*pesticides*) that cannot be found in the list of predicted pesticide names (*negative*). The true negative class (*tn*) consists of all words that are not organophosphorus pesticides and were correctly identified as such. Below, we describe the metrics to evaluate the methods studied in this work:

Precision (P): representing the correct keyword prediction among all predicted keywords.

$$\frac{tp}{tp + fp} \quad (2)$$

Recall (R): representing the correct keywords prediction among all ground-truth keywords.

$$\frac{tp}{tp + fn} \quad (3)$$

F1-Score: weighted harmonic mean of precision and recall.

$$F1 = 2 \frac{R \times P}{R + P} \quad (4)$$

6 Results

In this section, we present the results of three experiments. The first is a comparative analysis of six keyword extraction methods. The second evaluates multilingual keyword extraction methods for finding the exact match in the target corpus for a given seed keyword from the source corpus; in this study, an ‘exact match’ refers to the expected translation of the pesticide name. The third experiment assesses the same methods in retrieving a list of pesticide names in the target corpus for a given seed keyword in the source corpus. Here, we treat this task as the retrieval of ‘non-exact matches’, meaning that a seed successfully retrieves an organophosphorus pesticide even when it is not the expected translation.

6.1 Comparative Analysis of Keyword Extraction Methods

Figure 7 presents a comparative analysis of the number of keywords retrieved by different keyword extraction methods for the ORPHEUS-en corpus. The Ground Truth bar represents the total number of expected keywords, serving as a reference for evaluating the extractors’ performance. Among the automated methods, Simple Maths retrieves the highest number of keywords, followed by TF-IDF and YAKE, both of which extract a similar quantity. TextRank, MultipartiteRank, and KeyBERT retrieve fewer keywords, with TextRank exhibiting the lowest count. Table 2 presents a comparative analysis of the keyword extraction methods based on three evaluation metrics: precision, recall, and F1-score. The Simple Maths method outperforms all others, achieving the highest precision, recall, and F1 (respectively, 0.99, 0.70, and 0.82). TF-IDF follows with a significantly lower recall but relatively high precision, 0.89, indicating its ability to identify relevant terms while struggling with recall. KeyBERT exhibits similar precision, 0.89, but lower recall, 0.06, and F1-score, 0.11, suggesting that while it selects precise terms, it does not retrieve enough relevant ones. YAKE, TextRank, and MultipartiteRank show consistently low performance across all metrics, with F1-scores not exceeding 0.05. These results indicate that, for this task, statistical, reference corpus-based approaches, such as Simple Maths, are more effective than graph- or embedding-based keyword extraction techniques.

In Portuguese, similar to the results with the English corpus, the Simple Maths method retrieved the highest number of keywords (excluding the ground truth) and achieved the best overall performance, with an F1 score of 0.74. It showed a strong balance between precision (0.87) and recall (0.65),

⁶<https://pubchem.ncbi.nlm.nih.gov/>

Table 3. Multilingual keyword retrieval for pesticides (exact match)

$\theta_t >$	Approach	$\theta_k > 0.3$			$\theta_k > 0.6$			$\theta_k > 0.9$		
		Precision	Recall	F1-Measure	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
0.3	Basic	0.0048	0.6807	0.0095	0.0166	0.6649	0.0299	0.2662	0.2310	0.2070
	+ finetuning	0.0009	0.8451	0.0018	0.0015	0.8451	0.0029	0.2037	0.4225	0.2414
	Contextual	0.9592	0.4563	0.6184	0.8776	0.4175	0.5658	0.7419	0.2233	0.3433
	+ finetuning	0.9592	0.4563	0.6184	0.9388	0.4466	0.6053	0.7179	0.2718	0.3944
0.6	Contextual	0.5714	0.2718	0.3684	0.5306	0.2524	0.3421	0.6818	0.1456	0.2400
	+ finetuning	0.5714	0.2718	0.3684	0.5714	0.2718	0.3684	0.6786	0.1845	0.2901
0.7	Contextual	0.2245	0.1068	0.1447	0.2143	0.0874	0.1241	0.3333	0.0194	0.0367
	+ finetuning	0.2245	0.1068	0.1447	0.2083	0.0971	0.1325	0.3333	0.0194	0.0367

Table 4. LLM for keyword retrieval for pesticides (exact match)

Method	Precision	Recall	F1-Measure
GPT 3.5	0.3662	0.2149	0.2708
+ RAG	0.2727	0.0992	0.1455
+ GraphRAG	0.3803	0.2231	0.2812
GPT 4omini	0.2817	0.1653	0.2083
+ RAG	0.2581	0.1322	0.1749
+ GraphRAG	0.4085	0.2397	0.3021

indicating its effectiveness in identifying relevant domain-specific terms. In contrast, TF-IDF exhibited high precision (0.82) but significantly lower recall (0.31), suggesting that while it selected accurate terms, it failed to capture many relevant ones. Other methods, including YAKE, TextRank, MultipartiteRank, and KeyBERT, underperformed in both recall and F1 scores, particularly struggling with the domain-specific and potentially underrepresented vocabulary in Portuguese.

Although it offers reasonable performance, in the subsequent experiments, the list of ground truth keywords, presented in Section 5.3, is used as seed keywords. For comparison purposes, the results for keyword extraction by using the same methods on the Brazilian Portuguese dataset are provided in the Appendix B.

6.2 Exact Match Keyword Extraction

In this experiment, we evaluate the three multilingual keyword extraction methods presented in Section 5.2, where given a seed keyword in English (i.e., in the source corpus), the task is to retrieve the exact match in Portuguese (i.e., in the target corpus). Table 3 presents the results for the basic and contextual methods as a function of θ_k and θ_t . These parameters control the retrieval based on word and sentence similarities, respectively, between embeddings from the source and target corpus. The best results are marked in bold. For the basic approach, we observe a relatively high recall for the word similarity thresholds of 0.3 and 0.6, providing 0.6807 and 0.6649, respectively. Note that, for these thresholds, the model achieves very low precision and F1 score. It means that although it can find a considerable number of correct pesticide names, the amount of false positives is high, indicating that the model makes several wrong predictions. A greater threshold can mitigate the problem as it would remove the number of non-pesticide words being considered as positive examples. This comes with the cost of decreasing the number of true positives, as can be

seen for the cases where θ_k is set to 0.9, as in the last three columns of the first row. As a result, the recall drops to 0.2310, while the precision and F1 score increase to 0.2662 and 0.2070, respectively. Note that a similar trend is found when the basic approach is tested with the fine-tuned word embeddings. In such cases, recall and F1 score increase to 0.4225 and 0.2414, respectively, while the precision decays to 0.2037. In Table 3, we also present results for the contextual approach, where the sentence similarity is used and controlled by θ_t . We see that increasing the threshold eliminates the unrelated (or less similar) sentences. This approach proved to be effective in improving precision. The best results are achieved when both θ_t and θ_k are set to 0.3. In such a case, the achieved precision, recall, and F1 score are 0.9592, 0.4563, and 0.6184, respectively. This approach proved to be important to keep the model with high precision across different θ_k values. As we increase the word similarity to 0.6 and 0.9, the precision is still high for θ_t equals to 0.3 and 0.6. The precision only drops when θ_t is set to 0.7, with the reason being that many important words are left out because fewer relevant sentences are being retrieved from the target corpus. For this task of retrieving the exact match, the lowest thresholds seem to be the optimal solution combined with the contextual approach. Note that for low thresholds, the fine-tuning approach brings no benefits. However, it seems to be the best solution if the threshold is higher than 0.3, especially for the word similarity one.

Table 4 shows the results for the LLM and RAG approaches. The LLMs alone can achieve comparable performance with the basic solution. Nevertheless, by including external knowledge through RAG, results are improved, especially when using GraphRAG. The naive RAG is not effective in improving performance. We see that GPT 3.5 achieves a decent performance without the use of RAG, but it improves considerably with the use of GraphRAG. A similar trend is observed with GPT 4omini, with this solution achieving the best results. The superior performance of GraphRAG over naive RAG can be attributed to its structured knowledge organization. While naive RAG retrieves documents based on simple similarity, GraphRAG leverages semantic relationships between concepts. For example, when aligning the Portuguese keyword ‘*agrotóxico*’ (pesticide) with English, naive RAG might retrieve generic terms like ‘chemical’, whereas GraphRAG leverages its knowledge graph to correctly map it to domain-specific terms like ‘pesticide’ or ‘herbicide’, depending on context. This explains why

Table 5. Multilingual keyword retrieval for pesticides

$\theta_t >$		$\theta_k > 0.3$			$\theta_k > 0.6$			$\theta_k > 0.9$		
		Precision	Recall	F1-Measure	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
0.3	Basic	0.1462	0.9574	0.2369	0.1586	0.8894	0.2422	0.3358	0.3435	0.3200
	+ finetuning	0.0752	0.9986	0.1379	0.0809	0.9984	0.1465	0.4049	0.5151	0.4149
	Contextual	1.0000	0.4757	0.6447	0.9592	0.4563	0.6184	0.8065	0.2427	0.3731
	+ finetuning	1.0000	0.4757	0.6447	1.0000	0.4757	0.6447	0.7949	0.3010	0.4366
0.6	Contextual	1.0000	0.4757	0.6447	0.8571	0.4078	0.5526	0.8182	0.1748	0.2880
	+ finetuning	1.0000	0.4757	0.6447	1.0000	0.4757	0.6447	0.7500	0.2039	0.3206
0.7	Contextual	1.0000	0.4757	0.6447	0.7381	0.3010	0.4276	0.5000	0.0291	0.0550
	+ finetuning	1.0000	0.4757	0.6447	1.0000	0.4660	0.6358	0.5000	0.0291	0.0550

Table 6. LLM for keyword retrieval for pesticides

Method	Precision	Recall	F1-Measure
GPT 3.5	0.4225	0.2479	0.3125
+ RAG	0.6591	0.2397	0.3515
+ GraphRAG	0.4930	0.2893	0.3646
GPT 4omini	0.3239	0.1901	0.2396
+ RAG	0.8548	0.4380	0.5792
+ GraphRAG	0.4648	0.2727	0.3438

GraphRAG achieves higher precision (0.4085 vs 0.2581 for GPT-4omini) by reducing irrelevant retrievals.

The difference between GPT-3.5 and GPT-4omini’s performance (0.2812 vs 0.3021 F1 with GraphRAG) suggests model capacity plays a significant role. GPT-4omini’s larger parameter count enables better utilization of the retrieved context. However, our experiments show that even GPT-3.5 benefits from GraphRAG (F1 improvement from 0.2708 to 0.2812), indicating that the method’s effectiveness isn’t solely dependent on model size. The key factor appears to be the synergy between the model’s parametric knowledge and GraphRAG’s structured retrieval, where larger models can better exploit the high-quality context provided by the knowledge graph.

For instance, in Portuguese agricultural texts, the term ‘*defensivo*’ can ambiguously refer to ‘pesticide’ or ‘crop protectant’. GraphRAG’s knowledge graph helps disambiguate this by linking it to related terms (e.g., ‘fungicide’ or ‘insecticide’), while naive RAG might produce inconsistent alignments. The synergy between GraphRAG’s structured retrieval and the LLM’s parametric knowledge ensures more accurate bilingual keyword mapping, even when training data is limited.

6.3 List of Pesticide Keywords Extraction

In this experiment, the task is to retrieve a list of pesticide names from the target corpus, given a seed keyword from the source corpus (i.e., in English). There is no requirement to retrieve the exact match as, in some cases, practitioners or researchers are only interested in listing existing pesticide names in the target corpus. The performance of the models is presented in Table 5 and 6. As expected, results are typically higher than those achieved for the exact match task, which is more challenging. For the basic approach, recall is still high at the expense of precision. By increasing the word similarity threshold, the precision and F1 score also increase. For θ_k of 0.9, we attained 0.3358, 0.3435, and 0.3200 for precision, recall, and F1 score, respectively.

The finetuning helps to improve all three measures for the same threshold. However, it is less effective for θ_k less than 0.9. For those cases, it helps to improve recall, which means that it increases true positives and false positives as well. Thus, we can conclude that the best case scenario for the basic approach is to combine fine-tuning with a high threshold. For the contextual approach, we can see that the word similarity based on θ_k greater than 0.3 offers no different results for different θ_t nor the use of fine-tuning. Although the precision is high, which means that the model can predict true positives well, the recall is low, which means that the models provide many false negatives. For the θ_k greater than 0.3, results were not marked in bold as many of them were the same. The reason resides in the fact that while the sentence similarity threshold can restrict the context, it is not enough to impact the task of finding pesticide names that are part of the target list. In other words, the task is flexible enough to allow the method to find pesticide names no matter what context is given to it. The way to restrict the context is by choosing θ_k greater than 0.6 and 0.9. In this case, we see the effects of sentence similarity, which is controlled by θ_t . For θ_k greater than 0.6 and 0.9, there is a clear reduction in the precision, recall, and F1 score. The same trend is observed for θ_t greater than 0.6 and 0.7. Fine-tuning proved to be efficient in improving the results of the contextual approach in most of the scenarios.

Table 6 shows the results for the LLM and RAG approach for the retrieval of pesticide names without the requirement of an exact match. We observe that the LLMs can retrieve pesticide names, but their performance is improved with the use of RAG. Nevertheless, unlike the experiments with an exact match, the naive RAG was more effective in improving the performance of the LLM. Also, the best results are now from GPT 4omini and not from GPT 3.5. Some examples of ‘exact matches’ and ‘non-exact matches’ are briefly analyzed in Appendix D.

6.4 Linguistic Evaluation

The morphological structure of organophosphate pesticide names differs significantly from standard noun formation. Rather than simple root-plus-affix constructions, these names are built from mostly chemically motivated formants, each carrying semantic weight. This compositionality suggests that microtranslation should operate at the formant level to preserve both linguistic and chemical identity across

languages. For example, in the compound name ‘parathion’, *para-* means its molecule has substituents at the 1 and 4 positions of its benzene ring, whereas *-thion* means there is a carbon-phosphorus double bond. In Appendix C, we present the most common morphemes and substrings of pesticides’ names in both languages.

We also identified tensions between etymological accuracy and phonological adaptation, particularly in suffixes like *-thion*, *-ton*, and *-oxon*. In Brazilian Portuguese, phonological adjustments such as vowel insertion can obscure the underlying chemical function of the term. For instance, adding the vowel ‘a’ at the end of names such as ‘*etion*’, resulting in ‘*etiona*’, misleads its reader to believe there is a ketone (*cetona*, in Portuguese) group in the molecule, albeit there is not. Preserving etymologically consistent endings better supports cross-linguistic interpretability and aligns with ISO guidelines.

Orthographic conventions, such as the representation of voicing (s vs. ss) and gender assignment, further complicate translation. In the case of ‘glyphosate’, the element *-phos-* derives from ‘phosphonate’. In English, the ‘s’ is naturally devoiced in both ‘glyphosate’ and ‘phosphonate’. In Portuguese, however, these cases must be treated as archiphonemes, that is, the voicing of the phoneme represented by the grapheme ‘s’ depends on the following segment, as it appears at the end of the morpheme. Thus, in translations, while the ‘s’ in ‘*fosfonato*’ is devoiced due to the influence of the phoneme /t/, in ‘*glifosato*’ it becomes voiced under the influence of the vowel /a/. Furthermore, by adding ‘a’ to the end of a name, like the aforementioned ‘*etiona*’, its gender changes to feminine. In languages such as Portuguese, where gender agreement between nouns, pronouns, adjectives, and determiners is imperative, the variation in the gender of the word adds to the challenge of translation. Our data show that these features are often context-sensitive and influenced by both linguistic norms and domain-specific usage. Translators must balance phonetic naturalness with the need for terminological precision.

Finally, variation in translating formants like ‘methyl’ and ‘ethyl’ illustrates the challenges of lexical disambiguation in scientific domains. While multiple translations (e.g., ‘*metil*’ vs. ‘*metílico*’) coexist, our corpus suggests that more specific forms aid clarity, especially for non-expert readers. For example, although the standard common name of a molecule may be ‘*paration*’, researchers may also refer to it as ‘*etilparation*’ or ‘*paration etílico*’ depending on the context. These findings highlight how deep linguistic analysis can inform controlled vocabulary design and improve multilingual NLP for specialized domains.

7 Limitations

Since English is often overrepresented in the pre-training data, the tokenization method employed by multilingual BERT models may indeed be more finely tuned to English morphology and orthography, potentially leading to less optimal subword representations for less-represented languages like Portuguese. To address the tokenizer’s bias, we enhanced Portuguese representation by generating a large

synthetic parallel corpus via bidirectional translation using LLaMA-3.1-8B-Instruct. This bilingual dataset was then used to fine-tune a multilingual BERT model, improving its handling of Portuguese, especially for domain-specific terms. As shown in Table 3, this fine-tuning led to better alignment performance, with higher precision and recall in the keyword retrieval task.

8 Conclusion

This study highlights the importance of addressing terminological inconsistencies in pesticide-related chemistry. This work contributes to enhancing communication and regulatory standardization between high- and low-resource languages. By building corpora in Brazilian Portuguese and English, we assessed six keyword extraction methods and explored advanced techniques, including multilingual BERT embeddings and the RAG framework for keyword extraction. Our findings reveal that contextual sentence embeddings, followed by word embeddings derived from a fine-tuned BERT model, achieved the best performance when controlled by a similarity threshold. This research contributes to developing domain-specific multilingual tools, fostering improved cross-linguistic understanding in agricultural and chemical sciences.

Declarations

Funding

This research was funded, in part, by the São Paulo Research Foundation (FAPESP), processes 2021/08830-9 and 2019/14752-0, the National Council for Scientific and Technological Development (CNPq), process 130524/2021-2, and the Leading House for the Latin American Region (University of St. Gallen).

Authors’ Contributions

José Victor: Corpora Creation, Data curation, Validation, Formal analysis, Writing - final & original draft. **Hazem Amamou:** Coding & LLM experiments, Validation, Methodology - final setup, Data curation, Writing - final draft. **Rubing Chen:** Coding & fine-tuning experiments, Validation, Methodology - final setup, Data curation, Writing - final draft. **Elmira Salari:** Coding & Basic experiments, Validation, Writing - final draft. **Reto Gubbelman:** Methodology - original setup, Writing - original draft. **Christina Niklaus:** Methodology - original setup, Writing - original draft. **Talita Serpa:** Formal analysis, Writing - original draft. **Marcela Lima:** Formal analysis, Organophosphorus naming validation, Writing - original draft. **Paula Pinto:** Formal analysis, Project administration, Writing - original draft. **Shruti Kshirsagar:** Formal analysis, Writing - review & editing. **Alan Davoust:** Formal analysis, Writing - review & editing. **Siegfried Handschuh:** Project administration, Formal analysis, Writing - original draft. **Anderson Avila:** Project administration, Conceptualization, Methodology - final setup, Formal analysis, Writing - final draft.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and analyzed during the current study are available at <https://repositorio.unesp.br/entities/publication/715bd184-1c3f-441d-a514-8f0d8ffce15a>

Carbon Consumption Evaluation

To quantify the environmental impact of our computational experiments, we evaluated the carbon emissions associated with each phase of the research. Utilizing an NVIDIA A100 40GB GPU, the initial activities, including coding, debugging, bidirectional translation, dataset preparation, and BERT encoder training, amounted to approximately 30 machine hours, resulting in an estimated carbon emission of 3.24 kg eq. CO₂. Experiments involving the RAG model required approximately 14 minutes of GPU usage, resulting in emissions of 0.02 kg eq. CO₂. The GraphRAG model experiments entailed approximately 5.68 hours of GPU time, resulting in emissions of 0.53 kg eq. CO₂. These assessments highlight the carbon footprint of extensive computational tasks and underscore the importance of optimizing resource efficiency in research.

References

- Abakerli, R. B., Fay, E. F., Rembischevski, P., Vekic, A. M., Godoy, K., Maximiano, A. D. A., and Bonifácio, A. (2003). REGRAS PARA NOMENCLATURA DOS NOMES COMUNS DOS AGROTÓXICOS. *Pesticidas: Revista de Ecotoxicologia e Meio Ambiente*, 13(0). Number: 0. DOI: 10.5380/pes.v13i0.3162.
- Abulaish, M., Fazil, M., and Zaki, M. J. (2022). Domain-specific keyword extraction using joint modeling of local and global contextual semantics. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(4):1–30. DOI: 10.1145/3494560.
- Al-Rfou', R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics. Available at: <https://aclanthology.org/W13-3520>.
- Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*. DOI: 10.18653/v1/d19-1371.
- Bougouin, A., Boudin, F., and Daille, B. (2013). Topi-crank: Graph-based topic ranking for keyphrase extraction. In *International joint conference on natural language processing (IJCNLP)*, pages 543–551. Available at: <https://aclanthology.org/I13-1062/>.
- Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., and Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289. DOI: 10.1016/j.ins.2019.09.013.
- Cao, S., Kitaev, N., and Klein, D. (2020). Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*. DOI: 10.48550/arxiv.2002.03518.
- Carión, S. and Casacuberta, F. (2022). Few-shot regularization to tackle catastrophic forgetting in multilingual machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 188–199, Orlando, USA. Association for Machine Translation in the Americas. Available at: <https://aclanthology.org/2022.amta-research.14>.
- Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.findings-emnlp.261.
- Chambers, J. E. and Levi, P. E. (2013). *Organophosphates chemistry: fate, and effects*. Elsevier. Book.
- Delfino, R. T., Ribeiro, T. S., and Figueroa-Villar, J. D. (2009). Organophosphorus compounds as chemical warfare agents: a review. *Journal of the Brazilian Chemical Society*, 20:407–428. DOI: 10.1590/s0103-50532009000300003.
- Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., and Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*. DOI: 10.48550/arxiv.2404.16130.
- Finatto, M. J. (1996). Unidade e variação na língua portuguesa: a variação em terminologia. *Revista Internacional de Língua Portuguesa*, 15. Book.
- Finatto, M. J. B. and Kerschner, S. (1999). Dicionários especializados em tradução: cooperação entre o tradutor, o especialista e o terminólogo para a caracterização da terminologia e da linguagem da Química. *Cadernos do I.L. (dez. 1999)*, 21/22:p. 273–282. Available at: <http://hdl.handle.net/10183/247989>.
- Firoozeh, N., Nazarenko, A., Alizon, F., and Daille, B. (2020). Keyword extraction: Issues and methods. *Natural Language Engineering*, 26(3):259–291. DOI: 10.1017/s1351324919000457.
- Geary, R. (1953). Systemic insecticides, development of organic phosphates as systemic insecticides. *Journal of Agricultural and Food Chemistry*, 1(14):880–882. DOI: 10.1021/jf60014a003.
- Grootendorst, M. (2020). Keybert: Minimal keyword extraction with bert. Available at: <https://maartengr.github.io/KeyBERT/>.
- Hämmerl, K., Libovický, J., and Fraser, A. (2024). Understanding cross-lingual alignment—a survey. *arXiv preprint arXiv:2404.06228*. DOI: 10.18653/v1/2024.findings-acl.649.
- Hashemzadeh, B. and Abdolrazzagah-Nezhad, M. (2020). Improving keyword extraction in multilingual texts. *International Journal of Electrical & Computer Engineering (2088-8708)*, 10(6). DOI: 10.11591/ijece.v10i6.pp5909-5916.
- Hulth, A., Karlgren, J., Jonsson, A., Boström, H., and Asker, L. (2001). Automatic keyword extraction using domain

- knowledge. In *Computational Linguistics and Intelligent Text Processing: Second International Conference, CILing 2001 Mexico City, Mexico, February 18–24, 2001 Proceedings 2*, pages 472–482. Springer. DOI: 10.1007/3-540-44686-9_47.
- ISO (1981). ISO 1750:1981 Pesticides and other agrochemicals — Common names. Available at: [url:https://www.iso.org/standard/6370.html](https://www.iso.org/standard/6370.html).
- ISO (2018). ISO 257:2018 Pesticides and other agrochemicals — Principles for the selection of common names. Available at: <https://www.iso.org/standard/67998.html>.
- Kamrud, G., Wilson, W. W., and Bullock, D. W. (2022). Logistics competition between the u.s. and brazil for soybean shipments to china: An optimized monte carlo simulation approach. *Journal of Commodity Markets*, page 100290. DOI: 10.1016/j.jcomm.2022.100290.
- Kilgarriff, A. (2009). Simple maths for keywords. In Mahlberg, M., González-Díaz, V., and Smith, C., editors, *Proceedings of Corpus Linguistics Conference 2009*, University of Liverpool, UK. Available at: <https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf>.
- Kloske, M. and Witkiewicz, Z. (2019). Novichoks—the a group of organophosphorus chemical warfare agents. *Chemosphere*, 221:672–682. DOI: 10.1016/j.chemosphere.2019.01.054.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. DOI: 10.1093/bioinformatics/btz682.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474. DOI: 10.48550/arxiv.2005.11401.
- Lopes-Ferreira, M., Maleski, A. L. A., Balan-Lima, L., Bernardo, J. T. G., Hipolito, L. M., Seni-Silva, A. C., Batista-Filho, J., Falcao, M. A. P., and Lima, C. (2022). Impact of pesticides on human health in the last six years in brazil. *International journal of environmental research and public health*, 19(6):3198. DOI: 10.3390/ijerph19063198.
- Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., and Chi, Y. (2017). Deep keyphrase generation. *arXiv preprint arXiv:1704.06879*. DOI: 10.18653/v1/p17-1054.
- Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411. Available at: <https://aclanthology.org/W04-3252/>.
- Moncau, G. (2023). Conheça Adilson Paschoal, criador do termo ‘agrotóxico’ e parceiro de Ana Primavesi. *Brasil de Fato*. Available at: <https://www.brasildefato.com.br/2023/10/03/conheca-adilson-paschoal-criador-do-termo-agrotoxico-e-parceiro-de-ana-primavesi/>
- Section: Geral.
- Och, F. J., Tillmann, C., and Ney, H. (1999). Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Available at: <https://aclanthology.org/W99-0604>.
- Pinto, P. T. and Lima, M. d. F. (2018). A tradução na área de química orgânica: da adaptação à tradução literal. *Estudos Linguísticos (São Paulo. 1978)*, 47(2):573–585. DOI: 10.21165/el.v47i2.2050.
- Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*. DOI: 10.18653/v1/p19-1493.
- Piskorski, J., Stefanovitch, N., Jacquet, G., and Podavini, A. (2021). Exploring linguistically-lightweight keyword extraction techniques for indexing news articles in a multilingual set-up. In *Proceedings of the EACL Hackashop on news media content analysis and automated report generation*, pages 35–44. Available at: <https://aclanthology.org/2021.hackashop-1.6/>.
- Ragnarsdottir, K. V. (2000). Environmental fate and toxicology of organophosphate pesticides. *Journal of the Geological Society*, 157(4):859–876. DOI: 10.1144/jgs.157.4.859.
- Ruder, S., Vulić, I., and Søgaard, A. (2021). Hybrid approaches for low-resource word alignment. *arXiv preprint arXiv:2105.04556*. Available at: <https://arxiv.org/abs/2105.04556>.
- Sammet, J. and Krestel, R. (2023). Domain-specific keyword extraction using bert. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 659–665. Available at: <https://aclanthology.org/2023.ldk-1.72/>.
- Schuster, T., Ram, O., and Barzilay, R. (2019a). Fine-tuning pretrained language models for domain-specific tasks. *arXiv preprint arXiv:1909.00164*. Available at: <https://arxiv.org/abs/1909.00164>.
- Schuster, T., Ram, O., Barzilay, R., and Globerson, A. (2019b). Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1162.
- Sharma, P. and Li, Y. (2019). Self-supervised contextual keyword and keyphrase retrieval with self-labelling. DOI: 10.20944/preprints201908.0073.v1.
- Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., and Furtado, V. (2023). Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 268–282, Cham. Springer Nature Switzerland. DOI: 10.1007/978-3-031-45392-2_18.
- Soudani, H., Kanoulas, E., and Hasibi, F. (2024). Fine tuning vs. retrieval augmented generation for less popular knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 12–22. DOI: 10.1145/3673791.3698415.
- Souza, J. V. (2023). A corpus-based organophosphorus pesti-

cide bilingual glossary in Portuguese and English: a focus on denominative variation. Master's thesis, Unesp. Available at: <http://hdl.handle.net/11449/244658>.

Souza, J. V., Pinto, P. T., and Lima, M. M. d. F. (2022). Malationa, malation ou malatiom? a variação denominativa no processo de criação de um glossário bilingue da área de química de pesticidas. *Acta Scientiarum. Language and Culture*, 44(11):e55894–e55894. DOI: 10.4025/actascilangcult.v44i1.55894.

Verma, Y., Jangra, A., Saha, S., Jatowt, A., and Roy, D. (2022). Maked: Multi-lingual automatic keyword extraction dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6170–6179. Available at: <https://aclanthology.org/2022.lrec-1.664/>.

Wan, X. and Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. In *AAAI*, volume 8, pages 855–860. DOI: 10.5555/1620163.1620205.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., and Nevill-Manning, C. G. (1999). Kea: Practical automatic keyphrase extraction. In *Proceedings of the fourth ACM conference on Digital libraries*, pages 254–255. DOI: 10.48550/arxiv.cs/9902007.

Xia, P., Zhang, L., and Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information sciences*, 307:39–52. DOI: 10.1016/j.ins.2015.02.024.

Zhu, X., Li, L., Zhang, C., and Li, F. (2021). Domain-adapted word embeddings for improved sentiment analysis. *arXiv preprint arXiv:2103.06407*. Available at: <https://arxiv.org/abs/2103.06407>.

A Ground Truth

A.1 English

146 terms extracted from ORPHEUS-en

acephate	akton	aspon
azamethiphos-oxon	azinphos methyl	azinphos-methyl
azinphosmethyl	bensulide	bromophos
bromophos-ethyl	butonate	cadusafos
carbophenothion	chlorethoxyfos	chlorethoxyphos
chlorfenvinphos	chlorfenvinphos-oxon	chlorpirifos
chlorpyrifos	chlorpyrifos oxon	chlorpyrifos-methyl
chlorpyrifos-oxon	chlorpyrifos	chlorpyrifos me
chlorpyrifos methyl	coumaphos	crotoxyphos
crufomate	cyanophos	cythioate
DDVP	DEF	demeton
demeton S	demeton-methyl	demeton-O
demeton-S	demeton-S-methyl	dialifor
diamidfos	diazinon	diazinon-oxon
diazoxon	dicapthon	dichlofenthion
dichlorvos	dichrotophos	dicrotophos
dimethoate	dioxathion	disulfoton
ebufos	EPN	ethephon
ethion	ethoprop	ethoprophos
ethyl paraoxon	ethyl parathion	ethylparaoxon
ethylparathion	etrimfos	famfos
famphos	famphur	fenamiphos
fenchlorfos	fenchlorphos	fenitrooxon
fenitrothion	fensulfothion	fenthion
fonofos	formothion	glufosinate
glyphosate	heptenophos	iprobentfos
isazophos	isoazinphos	isodiazinon
isofenchlorphos	isofenitrothion	isofenphos
isomalathion	isophenphos	leptophos
malaoxon	malathion	methamidaphos
methamidofos	methamidophos	methidathion
methyl chlorpyrifos	methyl paraoxon	methyl parathion
methyl-paraoxon	methyl-parathion	methylchlorpyrifos
methylparaoxon	methylparathion	mevinphos
mipafos	monocrotophos	naled
oxydemeton methyl	oxydemeton methyl	oxydemeton-methyl
paraoxon	parathion	parathion-ethyl
parathion-methyl	parathionmethyl	phorate
phosalone	phosfolan	phosfon
phosmet	phosphamidon	phostebupirim
phoxim	pirimiphos ethyl	pirimiphos methyl
pirimiphos-ethyl	pirimiphos-methyl	pirimiphos me
pirimiphos methyl	profenofos	propetamphos
prothiofos	quinalphos	ronnel
sulfotep	sulfotep	sulprofos
tebupirimfos	tebupirimphos	temephos
TEPP	terbufos	tetrachlorvinphos
thiometon	thiophos	trans-phosphamidon
tribufos	trichlorfon	

A.2 Brazilian Portuguese

200 terms extracted from ORCHEUS-ptbr		
acefato	azametifós	azinfos
azinfós	azinfos metil	azinfós metil
azinfós-etílico	azinfos-metil	azinfós-metilico
azinfós-metilico	azinfosmetil	bromofós
bromofós etílico	bromofós-etílico	bromofós-metilico
bromofóssetílico	cadusafós	carbofenotion
carbofenotiona	carbofention	chlorpyrifós
clorfenrifós	clorfenvinfos	clorfenvinfós
clorpirifós	clorpirifós	clorpirifós etil
clorpirifós etílico	clorpirifós metilico	clorpirifós-metil
clorpirifós-metilico	clorpirifós-oxon	clorpirifós-oxon
clorpiripós	clorpirofós	clorhion
clortiofós	clortion	crufomato
cumafós	DDVP	DEF
demeton-S-metilico	demeton	demeton-S
demeton-S-metilico	diazinon	diazinon
diazinona	diazoxon	dichlorvos
diclorvos	diclorvós	diclorvos
dicrotofos	dimetoato	dimixion
dioxation	dissulfotom	dissulfoton
dissulfotona	disulfoton	edifenfós
etefom	etefon	ethion
etil paration	etil-paration	etion
etiona	etoprofos	etoprofos
etrinfós	etrinfós	etropofós
fenamifós	fenamifós	fenclorfós
fenitrothion	fenitrotion	fenitrotiona
fensulfotíon	fenthion	fention
fentiona	fentoato	forato
formotion	formotiona	fosalona
fosetil	fosetyl al	fosfamidom
fosfamidon	fosfamidona	fosmete
fostiazato	foxim	glifosate
glifosato	glufosinato	glyphosate
heptenofós	hostathion	iodofenfós
iprobefós	isazofós	isocarbophos
isomalathion	isomalation	isoxation
leptofós	malaixon	malaixona
malathion	malatião	malation
malationa	metamidofós	metamidofós
metaminofós	methamidofós	methamidophos
methyl parathion	metidation	metidationa
metil paraoxon	metil paration	metil paroxon
metil-paraoxon	metil-paration	metilparaoxon
metilparation	mevinfós	mevinfós
mipafox	monocrotofos	monocrotofos
naled	naled	paraoxon
paraoxon etílico	paraoxon-etílico	paraoxon-metilico
paraoxona	paraoxona etílica	parathion
parathion methyl	paratião	paratião-metil
paratiom metílico	paration	paration etílico
paration metílico	parationa	parationa metílica
parationa-etílica	parationa-metílica	parationametílica
paraxon	paroxon	pirazofós
piridafentiona	pirimifós metilico	pirimifós metilico
pirimifós-etílico	pirimifós-metil	pirimifós-metilico
profenofós	prothiofós	protiofós
quinafós	quinalfós	quinalfós
quinofós	quinolphos	sulfotepp
sulprofós	tebupirifós	tebupirinfós
TEEP	temefós	temephos
TEPP	terbufós	terbufós
tetrachlorvinfós	tiometon	tiometona
tolclofos metil	tolclofosmetil	triazofós
triazofós	tribufós	trichlorfon
triclorfom	triclorfon	vamidationa
vamidotion	vamidotiona	

B Keyword extraction in Portuguese

C Morphological approach

We conducted a parallel experiment to evaluate the performance of a cross-lingual morphological approach to retrieve keywords. The most common substrings were extracted from the pesticide names featured in the English ground truth and used as seeds for retrieving pesticide names in Portuguese. Specifically, we extracted the top 10 most frequent substrings for each substring length ranging from 2 to 5 characters, resulting in a final list of 40 substrings (Figure 8a).

These substrings include both complete morphemes (e.g., ‘chlor’, ‘thion’, ‘phos’, ‘met’, ‘para’, ‘oxon’) and truncated syllabic sequences (e.g., ‘pho’, ‘ophos’, ‘thi’, ‘hlorp’). To assess their retrieval potential, we identified all terms in the English corpus that contained these substrings and compared them with the ground truth. As shown in Table 7, there is a clear trade-off between recall and precision: using a larger set of substrings improves recall (up to 0.636 with 40 substrings), while reducing the list to 10 improves precision (0.014), yielding the highest F1 score of 0.027.

To extend this approach cross-lingually, we applied the substrings identified in English to the Portuguese corpus. We followed the translation guidelines proposed by Abakerli *et al.* [2003], who provide systematic rules for adapting pesticide names into Portuguese. For example, ‘ph’ becomes ‘f’, ‘th’ becomes ‘t’, and ‘y’ becomes ‘i’. Based on these guidelines, we translated the 40 English substrings into morphologically equivalent Portuguese substrings (Figure 8b).

Using these translated substrings as seeds, we extracted candidate pesticide names from the Portuguese corpus and evaluated their performance against the Portuguese ground truth. As reported in Table 7, the same precision-recall trade-off is observed. Notably, the Portuguese retrieval achieved slightly better results than English: a recall of 0.775 with all 40 substrings and a precision of 0.072 when using only 10, resulting in a maximum F1 score of 0.119. These findings suggest that cross-lingual morphological alignment can achieve high recall rates, but falls short in precision when compared to the contextual approaches presented in section 6.

D Case Study

In this case study, we present examples illustrating the application of Retrieval-Augmented Generation (RAG) for word alignment between English and Portuguese within the domain of pesticide-related chemistry. Using the gpt-3.5-turbo model, we aim to retrieve one and only one keyword from the target corpus based on the input keyword from the source corpus. Table 8 showcases several instances where the model aligns English pesticide names with their Portuguese counterparts or the most relevant terms available in the target context. Each example includes the input keyword, source context, target context, and the model’s response. In lines (1) and (5), the model successfully retrieved the expected exact translation of the input term. In samples (2) and (4), the outputs illustrate non-exact matches – that is, while the

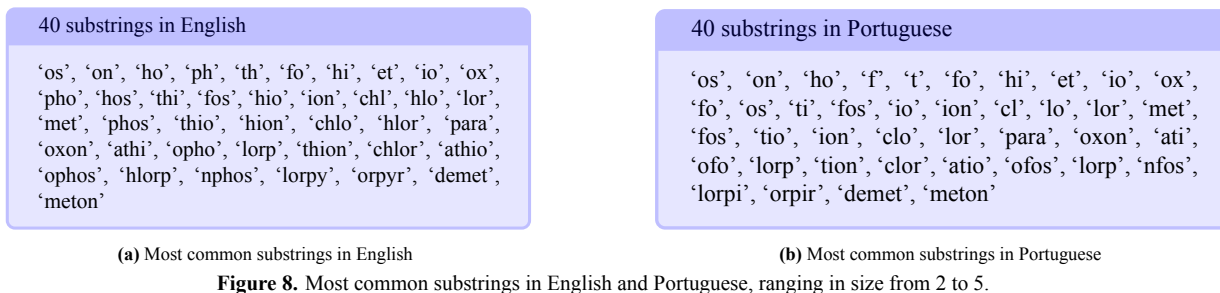


Figure 8. Most common substrings in English and Portuguese, ranging in size from 2 to 5.

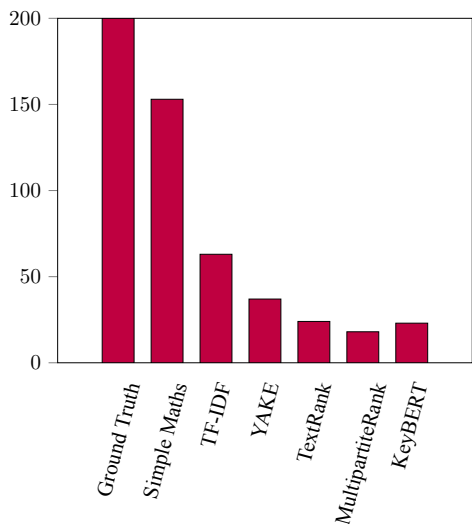


Figure 9. Number of correct keywords retrieved by each keyword extractor for the ORCHEUS-ptbr corpus

exact translation was not retrieved, the model still returned an organophosphorus pesticide. In contrast, line (3) represents a false positive: the retrieved term, although a pesticide (*metomil*), is neither the correct translation nor an organophosphorus compound.

Substrings	English			Portuguese		
	Precision	Recall	F1	Precision	Recall	F1
40	0.002	0.636	0.005	0.006	0.775	0.012
30	0.004	0.541	0.008	0.010	0.630	0.021
20	0.007	0.473	0.013	0.037	0.510	0.070
10	0.014	0.301	0.027	0.072	0.335	0.119

Table 7. Effect of substring number on pesticide name retrieval: English vs Portuguese

Table 8. Examples of RAG Application for Word Alignment Using gpt-3.5-turbo

Input Keyword	Source Context	Target Context	Model Response
(1) <i>bromophos</i>	Bromophos and bromophos-ethyl residues...	legislação brasileira.19 Alguns OPS – bromofós, diclorvós...	bromofós
(2) <i>azinphos-methyl</i>	Acephate Azinphos-methyl Bensulide Cadusafos...	INTRODUÇÃO azametifós DS NV traços...	Azametifós
(3) <i>carbophenothion</i>	O-Ethyl hydrogen isopropylphosphoramidothioate, O-ester with isopropyl...	cipermetrina (Cipermetrina®), metomil (Lannate®), clorpirifós (Lorsban®)...	Metomil
(4) <i>chlorfenvinphos</i>	chlorfenvinphos and bromfenvinphos in laboratory animals...	cipermetrina (Cipermetrina®), metomil (Lannate®), clorpirifós (Lorsban®)...	clorpirifós
(5) <i>mipafox</i>	Potential of Mipafox-Induced Delayed Neurotoxicity by Phenylfluoride (PMSF).	5) Mipafox: A) Solução estoque (0,1M)	mipafox