

EVOLVE: A SWISS-BRAZILIAN INTERDISCIPLINARY PROJECT  
TO RETRIEVE BILINGUAL TERMS OF PESTICIDE CHEMISTRY  
FROM TWO CORPORA IN ENGLISH AND PORTUGUESE

---

*EVOLVE: Projeto Interdisciplinar para Recuperar Termos Bilingües  
da Química de Pesticidas a Partir de Dois Corpora em Inglês e Português*

DOI: 10.14393/LL63-v39-2023-03

Paula Tavares Pinto<sup>\*</sup>

José Victor de Souza<sup>\*\*</sup>

Talita Serpa<sup>\*\*\*</sup>

Francine de Assis Silveira<sup>\*\*\*\*</sup>

Marcela Marques de Freitas Lima<sup>\*\*\*\*\*</sup>

Reto Gubelmann<sup>\*\*\*\*\*</sup>

Siegfried Handschuh<sup>\*\*\*\*\*</sup>

Christina Marianne Niklaus<sup>\*\*\*\*\*</sup>

---

<sup>\*</sup> Ph.D in Linguistics. São Paulo State University, Brazil. ORCID: 0000-0001-9783-2724. E-mail: paula.pinto(AT)unesp.br.

<sup>\*\*</sup> Master'sCandidate. São Paulo State University, Brazil. ORCID: 0000-0002-4587-4073. E-mail: jv.souza3(AT)gmail.com.

<sup>\*\*\*</sup> Ph.D in Linguistics. São Paulo State University, Brazil. ORCID: 0000-0003-3324-9593. E-mail: talita.serpa(AT)unesp.br.

<sup>\*\*\*\*</sup> Ph.D in Linguistics. Federal University of Uberlândia, Brazil. ORCID: 0000-0002-3962-3972. E-mail: francinesilveira(AT)ufu.br.

<sup>\*\*\*\*\*</sup> Ph.D in Chemistry. São Paulo State University, Brazil. ORCID: 0000-0003-3191-2494. E-mail: marcelamarques.lima(AT)unesp.br.

<sup>\*\*\*\*\*</sup> Ph.D in Computer Science. University of St. Gallen, Switzerland. ORCID: 0000-0001-6141-4168. E-mail: reto.gubelmann(AT)unisg.ch.

<sup>\*\*\*\*\*</sup> Ph.D in Computer Science. University of St. Gallen, Switzerland. ORCID: 0000-0002-6195-9034. E-mail: siegfried.handschuh(AT)unisg.ch.

<sup>\*\*\*\*\*</sup> Ph.D in Computer Science. University of St. Gallen, Switzerland. ORCID: 0000-0001-9344-8127. E-mail: christina.niklaus(AT)unisg.ch.

**ABSTRACT:** Brazil is the second-largest exporter of soybeans and the largest importer of pesticides worldwide. In Pesticide Chemistry, we frequently find terms that are coined in English and still do not have equivalents in Portuguese, such as the term *dichlorvos* which has variants in Portuguese, such as “*diclórvós*”, “*diclorvos*” and “*diclórvos*”. Based on terminological studies and corpus linguistics, this paper aims to present an interdisciplinary approach developed by a Swiss-Brazilian research group to produce a bilingual glossary of Pesticide Chemistry. To do so, we compiled two corpora of academic texts in English and in Portuguese to (i) find definitions of organophosphorus pesticides in English and their equivalents in Portuguese, (ii) retrieve definitions for the terms based on their contexts in concordance lines. As a result, we identified spelling differences in the terms in Portuguese which have been solved based on the guidelines of normalizing institutions from the area.

**KEYWORDS:** Corpus Linguistics. Translation. Terminological variants. Pesticide Chemistry. Internationalization.

**RESUMO:** O Brasil é o segundo maior exportador de soja e o maior importador de agrotóxicos do mundo. Na Química de Pesticidas, frequentemente encontramos termos em inglês que ainda não possuem equivalentes em português como *dichlorvos*, em inglês, que apresenta as variantes “*diclórvós*”, “*diclorvos*” e “*diclórvos*”, em português. Este problema leva a interpretações incorretas por pesquisadores e pela sociedade em geral. Com base na terminologia e na linguística de corpus, este artigo apresenta uma abordagem desenvolvida por um grupo suíço-brasileiro para produzir um glossário bilíngue de Química de Pesticidas. Para tanto, compilamos dois *corpora* de textos acadêmicos com a finalidade de: (i) encontrar definições de pesticidas organofosforados em inglês e seus potenciais equivalentes em português, (ii) recuperar definições dos termos a partir dos contextos nas linhas de concordância. Os resultados identificaram diferenças ortográficas nos termos em português que foram resolvidas com base nas diretrizes normatizadoras da área.

**PALAVRAS-CHAVE:** Linguística de Corpus. Variantes terminológicas. Tradução. Química de Pesticidas. Internacionalização.

---

## 1 Introduction

The Sustainable Development Goals (SDGs) “address the global challenges human beings face, including those related to poverty, inequality, climate change, environmental degradation, peace, and justice”. As a consequence, researchers of all countries have directed their studies to find solutions for the issues pointed out by the UN Agenda 2030. Given that, studies and publications involving the SDGs have emerged, especially the ones related to Health and Environment. In this sense, this paper will tackle the translation and terminological issues regarding publications in the area of Pesticide Chemistry in Brazil and how these linguistic issues can influence the health and safety of Brazilian communities and cities.

In 2016 and 2017, Brazil became the second-largest exporter of soybeans globally and has remained in top positions since then. As a consequence, it has become the largest importer

of pesticides in the world. Among the imported chemical products there is “glyphosate”, for example, which is the leading international herbicide in sales. Given this context, researchers of Chemistry seek to study and create a set of environmentally sustainable methodologies for the degradation of pesticides that are illegally traded or that are part of expired stocks, since they are highly harmful to the health of the general population. However, one problem researchers have found is the lack of terminological standardization of Pesticide Chemistry terms in English and in Portuguese, which has led to misinterpretation of product labels as well as wrong use of the terminology that should be accurately translated to avoid ambiguity not only by researchers but also by technicians and farmers who have to deal with those products. Taking that into consideration, this paper will tackle the translation and terminological issues regarding publications in the area of Pesticide Chemistry in Brazil and how these linguistic issues can influence Brazilian communities and cities on the right to healthy food.

Even though the need for studies of Pesticide Chemistry is internationally recognized, when it comes to publications in this area in Brazil or research papers that disseminate studies in Portuguese, the vocabulary is extremely based on historical terminologies of a complex language that is mostly written in English. One example that illustrates this issue is the use of the term malathion, which is a pesticide usually translated to Portuguese as i) malathion (identical to the English form), ii) malatiom (adapted, however not representative of the phosphoryl group), or most appropriately, iii) malation (evoking the correct chemical group) (SOUZA *et al.*, 2022). All these terms are used in labels and prescriptions without much regulation, which is the Portuguese’s right equivalent. Consequently, the same compound is often registered as three or even five different compounds in business or academic studies as well as master's theses and doctoral dissertations. Therefore, as it has been shown, there is still no consensus on the use of standardized terms in Portuguese to refer to their equivalents in English, which has led authors of scientific papers to use terms in English or neologisms instead of searching for the correct terminological equivalents in both languages.

Considering the terminological issues previously discussed, we started an interdisciplinary partnership among researchers of Translation, Linguistics, and Chemistry at the São Paulo State University (UNESP). This research team has received financial support from the São Paulo Research Foundation (FAPESP) and the National Council for Scientific and

Technological Development (CNPq). However, the project reached a point that needed specialized technical support in the selection procedures since the search has shown other distinct terms in Portuguese being used as a translation for single components in English. For this reason, we submitted a project in partnership with specialists in Large-Scale Text Mining and Distributional Semantics from the University of St. Gallen, Switzerland. The project is called “EVOLVE: language as a tool for EnVirOnmentaLly sustainable actions in deVEloping countries: for the right to healthy food and has been supported by the Swiss funding agency called Leading House.

The computational methods used by the Swiss researchers have been of major importance and relevance for the compilation of terms that will compose a digital glossary on Organophosphorus Pesticides proposed here.

This paper describes the advances taken with the association of a Corpus-based methodology with Natural Language Processing and Chemistry. Based on the previous context, we present the procedures for the compilation of a digital glossary of Chemistry, more specifically in the Organophosphorus Pesticides field, to be presented to academics, graduate and undergraduate students of Chemistry, as well as to translators, and people interested in this field.

Research Questions:

- (1) What are the most representative terms in the Organophosphorus Pesticides field in English and Portuguese?
- (2) How to prepare a bilingual descriptive digital glossary in a field with high denominational variation such as Pesticides Chemistry?
- (3) How can the data collected from our corpus contribute to the teaching, in English and Portuguese, of those terms?

To answer the questions presented here, this paper will be divided into the following sections: (i) Terminology and Translation in Chemistry, more specifically on Organophosphate Pesticides; (ii) Methodology involving corpus compilation and term extraction; (iii) Analyses regarding seed extractions, first computational trial and information on terminological form; (iv) discussion and (v) conclusion.

## 2 Terminology and Translation in Chemistry: Organophosphate Pesticides

Phosphorus was discovered in 1669, 350 years ago (SHARPLEY *et al.*, 2018) and it is an extremely controversial element nowadays (EMSLEY, 2011). However, the discovery and study of the first phosphorus-based organic compounds were only systematized in 1854, with the synthesis of tetraethyl pyrophosphate by Philippe de Clermont (COSTA, 2017). The study of organophosphorus substances has attracted the attention of the best contemporary researchers due to the discovery, at an exponential pace, of applications ranging from pesticides to the cure of the most offensive diseases of today (DEMKOWICZ *et al.*, 2016; COSTA, 2017).

It is in this context that material that facilitates English-Portuguese bilingual communication becomes essential in the scientific environment. Thus, it was observed, through the study, that there is a lack of reference materials related to the terminology of organophosphorus compounds, as well as a lack of consistency in the translation of the available terms into the Portuguese language, which greatly complicates the dissemination of knowledge in the sub-area of chemistry, especially concerning the development of basic research and chemistry students' access to consistent terminology.

On the other hand, informative works (or basic research) involving phosphorus chemistry are written with terminological inconsistencies or even with terms in English, since they are fed by teaching materials that have not received adequate terminological-translational treatment. The effects generated by the lack of standardization are serious: incorrect reading of pesticide leaflets, or even medicines, and difficulty in establishing a univocal discourse in the area, as established by IUPAC standards (McNAUGHT, 1997) and ISO standards.

The International Union of Pure and Applied Chemistry (IUPAC) has made an effort to make communication in this area of knowledge more transparent among nations. However, the terminology in Chemistry is still regional, unlike its symbology, which is universal. When new terms are proposed, through translations, they are poorly standardized and are not always followed by academic scholars. Azenha Jr. (1999) emphasizes cultural aspects in translated technical texts, as they are part of the relationship between language, culture, text, and translation. That margin leads to variation, which is accepted as an inherent part of specialized language, now observed from levels of specialization and, thus, with different levels of

description (CABRÉ, 1999a; CABRÉ, 1999b; CABRÉ *et al.*, 2007; CONDAMINES, 2010). For Bowker and Pearson (2002) and Serpa and Cerna-Chávez (2020), despite the fact that technical communication is more controlled, terminology is a dynamic area and allows subjective decision-making by its users.

According to Cabré (1999a) there is a clear interdisciplinarity between Translation and Terminology, yet very little has been studied about the characteristics and motivations for this relationship, and even less has been considered about the limits between them both. In Brazil, the language direction of translated texts has long been from English to Portuguese. However, with increased human interactions and international business exchange, the demand for translated texts into Portuguese has significantly increased. Translators have often created neologisms or even paraphrased terms “to accommodate semantic equivalences” (KRIEGER; FINATTO, 2004). On the other hand, Pinto and Lima (2018) highlight some of the problems specialists of Chemistry may find in published papers:

Chemistry researchers report that there is still no linearity in the choice of translation equivalents for each term and that, often, they use linguistic decal as a choice, which is understood here as an adaptation of the term in the target language, or the borrowing of a term in English even after an equivalent in the target language has already been proposed, in order to clarify the process on focus. (PINTO; LIMA, 2018, p. 574)

Therefore, we observe that Terminology tends to provide the necessary material for translation activity, especially concerning equivalents. Professionals in the field can count on quick access to use the correct terms from the most diverse fields of technical-scientific production.

Considering the terminological issues previously pointed out (PINTO; LIMA, 2008; SOUZA, 2019; SOUZA *et al.*, 2022) we have set an interdisciplinary collaboration to fill in the terminological gap in the area of Pesticide Chemistry by analyzing keyness scores through the use of Sketch Engine (KILGARIFF *et al.*, 2014) to identify terms by using the theoretical background of Corpus-Based Translation and Terminology (PAIVA *et al.*, 2008; CAMARGO *et al.*, 2012; KRIEGER; SANTIAGO, 2014) and Natural Language Processing (QASEMIZADEH; HANDSCHUH, 2014; GUBELMANN *et al.*, forthcoming). When variants were identified among the equivalent terms in Portuguese, they were classified according to their type. After that,

statistical math was applied, as well as specialists in pesticide chemistry were consulted to define the most suitable variants.

We have followed a corpus-based perspective towards Terminology and Translation to develop this study. In the following sections we present our point of view to analyze the terms in the area of Pesticide Chemistry.

## 2.1 Terms in this study

In Terminology, Andrade (2001) points out that terminological studies deal with “the term, that is, the specialized word, the concepts inherent to the various specialized subjects” (ANDRADE, 2001, p. 192). Specialized terms understood as the “designation, by means of a linguistic unit, of a concept defined in a language of specialty” (ISO 1087, 1990, p. 5 *apud* BARROS, 2004, p. 40) will be analyzed in this study.

According to Barros (2004, p. 105), the usage of a term within a distinct phrase is a decisive criterion since it considers the stability of the link between syntagmatic sequence and the unique meaning of the word. Users develop a strong semantic-syntactic understanding and memory of a phrase after repeated use in a syntagmatic sequence. As a result, the syntagma acquires shape and meaning stability. To the author, a simple term is built of a single radical, with or without affixes (ISO 1087, 1990, p. 7 *apud* BARROS, 2004, p. 40), whereas a complex term is constituted of two or more radicals, to which other components can be added (ISO 1057, 1990, p. 7 *apud* Barros, 2004, p. 40). Regarding the compound terms, Barros (2004) suggests that

[...] they are also lexical units formed by two or more radicals. However, they are distinguished from complex terms by the high degree of lexicalization and by the set of lexical and/or grammatical morphemes that constitute them, in a situation of non-autonomy graphically represented by the use of the hyphen. [...] It should be noted that we consider complex lexical units by agglutination (as nobleman, although, etc.) and by the hyphen-free juxtaposition of two or more radicals as simple terms.<sup>2</sup> (BARROS, 2004, p. 100)

---

<sup>2</sup> [...] também são unidades lexicais formadas por dois ou mais radicais. Distinguem-se, no entanto, dos termos complexos pelo alto grau de lexicalização e pelo conjunto de morfemas lexicais e/ou gramaticais que os constitui, em situação de não-autonomia representada graficamente pela utilização do hífen. [...] Cumpre ressaltar que consideramos as unidades lexicais complexas por aglutinação (como fidalgo, embora etc.) e pela justaposição sem hífen de dois ou mais radicais como termos simples.

In our research, the list of terms to be compiled as a glossary will follow Barros (2004) definition, to whom:

Glossary (tolerated term: bilingual dictionary, multilingual dictionary): can be located at both the system level and the standard(s). Its main characteristic is not to present definitions, but only a list of lexical or terminological units accompanied by their equivalents in other languages. (BARROS, 2004, p. 144)

The glossary to be presented in this study will be used to better inform Chemistry researchers, students and translators how Organophosphorus terms are used within their communicative contexts. To do so, we also follow the Cabré's (1999b) concepts of Communicative Theory of Terminology, since the author points out that, when translating to different professional areas, it is crucial to reach out for specialized terminology in order to successfully acquire communication and social connection across different specialists worldwide. Terminology can help those specialists by elaborating consultation materials to help this group.

Another aspect to emphasize is that term frequency is not evenly distributed. Sinclair (1991) and Halliday (1992, 1994) argue from a probabilistic perspective that language is a system of probabilities, the most evident of which is the frequency with which words are used. Berber Sardinha points out that:

[...] frequency is an inseparable attribute of the word, as it reveals its observed occurrence in use. The frequency of use (high, low, intermediate, etc.) has a defining role in the word, giving it a trait as inseparable as meaning. (BERBER SARDINHA, 2004, p. 162-163)

The observation of co-occurrences of two or more words are more related to specific terms than to others for the analysis of patterns in terminological languages.

## 2.2 Variation in Terminology

Although the assumptions of Terminology suggest that the search for translations of terms is linked to the purpose of equivalence and correspondence, variation seems to be a very common issue and needs to be considered when dealing with texts of Chemistry.



Taking that into consideration, “terms, in linguistic and social environments, are entities capable of variation” (FAULSTICH, 2002, p. 70). The author considers that it is possible to verify terminological variation since languages of specialty can vary in their form and content, in diachrony and synchrony, being possible to state that

(...) no language stage is a homogeneous block, although it is regular. Each language stage, in turn, is limited by complexes of linguistic varieties, which are intertwined by language impulse and tend to present: i) variation as a process; ii) variants as natural evolution protocols; iii) the change as a product of the change in communicative schemes. (FAULSTICH, 2002, p. 28)

Esteves (2010) follows the theoretical basis pointed out by Faulstich and notes that the concept of terminological variation delimits the definition of term and the understanding of its use within the different linguistic systems. In this researcher's perception, the function of a term is immersed in the conjuncture of the different languages of specialty. We also observed that this proposition fits the purposes of our research, since we seek to evaluate the possible divergences between the composition of the terms.

Thus, the functionality of a term is directly related to the context in which such terminologies are used within the various areas of specialty. The analysis by Faulstich (2001; 2002) and Esteves (2010) corroborate the ideas that terms assume specific functions “according to the context of use”; and that, under similar conditions of use, “they will be considered variants of each other” (FAULSTICH, 2002, p. 75).

Thus, the theorists also point out a series of postulates that guide this theory of variation, namely:

- a) dissociation between terminological structure and homogeneity or univocity or monoreferentiality, associating to the terminological structure the notion of ordered heterogeneity;
- b) abandoning the categorical isomorphism between term-concept-meaning;
- c) acceptance that, as terminology is a fact of the language, it accommodates variable elements;
- d) acceptance that the terminology varies and that this variation may indicate an ongoing change;
- e) terminology analysis in linguistic co-texts and in discursive contexts of written and oral languages. (FAULSTICH, 2002, p. 76)

With regard to the translation process, these factors motivate a correlation between possible changes in analytical perspectives from one language to another, by identifying the alternation of functions that the variants assume within the linguistic and social communities. In the conception of Faulstich (2002, p. 76), the terms are closely related to the position they exercise within a social and cultural system, their performance being part of an entity of a pragmatic and empirical nature, which conditions the possible "mechanisms of variation".

In the case of competitive variants, Faulstich (2002) states that

[...] are those that relate meanings between lexical items of different languages, that is, lexical items of a language B fill gaps in a language A. Competitive variants suffer, in their performance, intersections, due to the very foreign nature of the expression. This phenomenon occurs when the language structure of the foreign term is disturbed by vernacular language structures: the mixture of formants activates variation. (FAULSTICH, 2002, p. 77)

Thus, the author considers that the occurrence of this type of variants occurs through pairs of linguistic loans and vernacular forms and adds that

Linguistic loans are lexical items that originate from a foreign language and then, in the social context of the receiving language, become variants because they cause the emergence of an equivalent vernacular form, because of the linguistic environment foreign to its natural permanence. (FAULSTICH, 2002, p. 77)

Consequently, in the translation process of texts based on cultural terms, it is possible to understand that there are several stages in which the terminology is adapted to the numerous types of variations and loans. In the context of Translation, therefore, factors such as discourses, regionality, geography, temporality are important aspects to be considered during the activity of the professional in the area.

We consider, therefore, that a view based on the possibility of interaction between the terms will serve us to contextualize the analysis of the data and the teaching and learning of the translation activity; as well as to verify how the relationship between Terminology and Translation can influence the changes of concepts and settings in a language of specialty focused on Organophosphates Pesticides.

In the following section we will present the methodology used in this study so far.

### 3 Methodology

By using computerized tools (SARDINHA, 2000; KILGARRIFF *et al.*, 2014), a specialized bilingual corpus was compiled, from which relevant terms have been extracted. The corpus compilation in this research is based on Tognini-Bonelli's proposal (2001), and on the studies of Paiva *et al.* (2008) and Paiva (2009). Tognini-Bonelli considers important the use of comparable corpora, that is to say, one corpus with texts originally written in language 1 (L1, in the case of the present work, English) and another one with texts originally written in language 2 (L2, in this case, Portuguese), because they allow a better identification of the form and function of words.

According to Tognini-Bonelli there will be stages to be followed when finding equivalent terms in corpora. The first stage is identification and classification of the lexical and grammatical patterns within the context of a word or expression. The second stage is the first sense recognition (*prima-facie*) of the word, comparing form and function between L1 and L2. The third stage considers function as the observation of the form of realization (collocational or colligational pattern) in L2. This method suggested by Tognini-Bonelli (2001) is related to the process of decoding and encoding in another language. With the help of corpora in two languages, the researcher has access to the term as it is used, in L1 and L2, within a specific context, which enables a more adequate choice of the corresponding term for translation, based on actual evidence of use in both languages.

#### 3.1 Corpus compilation

Two comparable corpora were simultaneously created and compiled (tagged) for the present study. The first one, in Portuguese, is named ORCHEUS (Organophosphorus Chemistry Corpus). It consists of 84 academic texts that address environmental and health issues surrounding pesticide poisoning. Those texts were published between 1996 and 2020 in peer-reviewed scientific journals or thesis/dissertation banks and, for the purpose of this study, were classified according to their characteristics into the categories of article, thesis, and dissertation. Each text had some excerpts that were mainly in English removed (such as the abstract and bibliographic references) to improve keyword-list generation results in Portuguese. Finally, we gathered a corpus whose number of tokens (all word occurrences) is

830,144 while the number of words or types (repeated occurrences counted as one) is 621,213. This proximity between token and word numbers indicates rich lexical diversity.

The second corpus, in English, was named ORPHEUS (Organophosphorus and Phosphorus Chemistry Corpus) and consists of 201 academic texts published between 1943 and 2022, with a total of 3,128,199 tokens and 1,988,863 words. Just as ORCHEUS, its documents are papers published in academic journals or dissertation banks and topics also revolve around organophosphorus compounds. However, texts concerned with phosphorus and its inorganic derivatives (without carbon and oxygen) were also added to this corpus to cover other substance names that, because of the lack of research interest in them in Brazil, would not be covered in our corpus in Portuguese. Furthermore, this topic widening allows us to look into a range of phosphorus functions (both organic and inorganic), which may also vary, as found by Rocha, Lima, and Serpa (2020). The same procedure, however, could not be performed with ORCHEUS since scientific production in phosphorus chemistry in Portuguese is not as numerous as in English. This difference in approach resulted in one corpus being larger and more comprehensive than the other; this apparent problem, however, does not compromise our goals as ORPHEUS's wider scope allows us to find missing terms in ORCHEUS.

With the aim of generating a list of seed words to be part of terms, we used a list of Keywords that contrasts a specialized corpus to a general corpus in the same language, which was English ten ten. As explained by SOUZA *et al.* (2022), Sketch Engine uses a method called simple math (KILGARRIFF, 2009), which generates a keyness score that ranks the words in the corpus based on the following formula:

$$\frac{fpm_{focus} + N}{fpm_{ref} + N}$$

In this case,  $fpm_{focus}$  represents the relative frequency (per million) of a given word in the specialized corpus, and  $fpm_{ref}$  represents the relative frequency (per million) of a given word in the reference corpus. The relative frequency is represented by the formula below:

$$fpm = \frac{\text{number of hits} \cdot 1.000.000}{\text{corpus size}}$$

In the previous formula, for a given word in a corpus, whether the study or reference, there is a given number of hits (frequency of the word in the corpus) which is normalized in parts per million and divided by the corpus size (the total number of tokens, or items, from the corpus in question).

In the following section we will present some results from the combined methodology we have been using between Corpus Linguistics to generate a list of terms, in English, related to pesticides and, next, how this list is being used by Computational specialists to help the search for equivalent terms in Portuguese.

#### **4 Analyses**

This section will be divided into three parts which will show how we generated a list of seed-words to point out simple terms, the first trial for finding terms in English and their equivalents in Portuguese and a tentative terminological form to be used in the compilation of a digital glossary.

##### **4.1 Generating Seeds from the corpus**

Once we had selected the simple terms generated with the simple math, as previously explained, we produced a list of simple terms in English to be shown to the specialist in Chemistry, who indicated the organophosphorus pesticides. The list was kept in a spreadsheet as we can see below:

Table 1: first 15 seed words in English from ORPHEUS corpus

1	Item	Concept
2	chlorfenvinphos	pesticide
3	crotoxyphos	pesticide
4	methyl-paraoxon	pesticide
5	methylparaoxon	pesticide
6	ethylparaoxon	pesticide
7	paraoxon	pesticide
8	ddvp	pesticide
9	dichlorvos	pesticide
10	bidrin	pesticide
11	dichrotophos	pesticide
12	dicrotophos	pesticide
13	mevinphos	pesticide
14	phosdrin	pesticide
15	azodrin	pesticide

Source: the authors.

In total there were 64 pesticides which were presented to the Swiss group in order to apply their own methodology for finding possible equivalent terms in Portuguese.

#### ***4.2 First computational trial to find equivalents in English and in Portuguese***

The list of organophosphorus pesticides in English was used to be contrasted with ORCHEUS corpus, in Portuguese, with a new methodology developed by the Swiss group (GUBELMANN *et al.*, forthcoming). The first result was the list presented in Table 2, where we can see: (i) an English term (Sentence-key), (ii) an equivalent candidate in Portuguese (Candidate), (iii) the measurement of cosine similarity between sentence-embeddings in English and in Portuguese (Cand-Cosine) and (iv) a sentence in the target corpus where the candidate term could be present (Sentence).

This list was presented to the linguists from the Brazilian group, as well as to the specialist in Chemistry, who discussed the results produced in this first trial. Although the list of possible equivalents in Portuguese was not precise, the contexts in the sentences were always related to the pesticide it was linked to, which was considered a positive indication of the method so far. At the moment, the group has been working on new trials, not only between English and Portuguese, but in Spanish as well, since the similarity between these last languages may become an important hint to improve this methodology.

Table 2 – seed words in English from ORPHEUS and possible equivalents in Portuguese based on contexts from ORCHEUS.

Sentence-Key	Candidate	Cand-Cosine	Sentence
azinphosmethyl	phosphamidon	1,094672322	phosphamidon, monocrotophos, A frequência de SCE também foi significativamente maior no grupo quinalphos, dimetoato, exposto em todas as durações de exposição (1 a 10 anos, 11 a 20 anos e fenvelrate ou cipermetrina >20 anos).
demeton-methyl	methylazinphos	1,094556451	12 inseticidas organofosforados, Foram investigados os efeitos de 3 combinações de trichlorfon com cada derivado de metil, malathion, parationa-metílica e methylazinphos.
glyphosate	Glyphosate	1,094305277	Keywords: Glyphosate.
oxydemeton-methyl	clorpirifos-oxon	1,094272614	Kralj <i>et al.</i> , (2007) em um estudo de produtos de degradação de pesticidas organofosforados relata que, para o pesticida clorpirifos, um dos produtos encontrado foi o análogo clorpirifos-oxon.
methamidophos	methamidophos	1,094191313	Exposure to methamidophos at adulthood adversely affects serotonergic biomarkers in the mouse brain.
azinphos-methyl	clorfenvinfos	1,093927383	O objetivo do presente estudo foi desenvolver e validar um método analítico, empregando a dispersão da matriz em fase sólida (DMFS), seguida pela cromatografia gasosa acoplada à espectrometria de massas (GC-MS), na análise de resíduos dos pesticidas clorfenvinfos, fipronil e cipermetrina, os quais são aplicados no rebanho bovino no combate ao carrapato <i>Boophilus microplus</i> , utilizando como matriz de estudo o plasma bovino.
oxydemeton-methyl	Clorpirifos-oxon	1,093624115	91 A) B) Figura 50: Espectros de massa dos compostos separados por CG-MS: A) Pesticida clorpirifos
oxydemeton-methyl	azinfos-metílico	1,093557715	Os agrotóxicos detectados são: heptenofos, clortiofos, PBO (piperonylbutoxide), dieldrina, azinfos-metílico, dodecacloro, parationa-etílica e monocrotofos.
methylparathion	methylbromphenvinphos	1,093243122	A cinética de metilação de bases purínicas no DNA pelo parationa-metílica e [14C]malathion mostrou que esse processo foi um pouco lento, atingindo seu máximo methylbromphenvinphos após 96 horas da reação.

Source: elaborated by the authors.

### 4.3 Information on the terminological record

In this part of the paper, we present our first model of terminological record ; it has been discussed so far in Souza's study (forthcoming) and the Brazilian team with the help of a terminologist. This record is a means for recording, in a structured set of fields, the terminological data for a specialized concept. It is important to point out that the variants of the selected terms, as well as the contexts in which they are found, are crucial to this study since we are taking both into account before proposing the best equivalent in Portuguese.

Table 3 – Part 1 of the terminological card.

English			
Number: 1	Entry: malathion	Classifying symbol: 1.1.1.2.1	PoS: substantivo
			Type: simples
Variant I:			Notes:
Context 1:  In addition, adsorption of <b>malathion</b> was lower in soils whose organic matter was destroyed by heating than in control soils. <dissen03>  These issues can be demonstrated on the example of <b>malathion</b> , a very common pesticide. <dissen04>			
Definition:			
IUPAC name: diethyl 2-[(dimethoxyphosphorothioyl)sulfanyl]butanedioate			

Source: the authors.

In the first part of the card, readers will find: (i) the number of term in the list; (ii) the entry term in English; (iii) its classifying symbol; (iv) a variant if there is one; (v) the part-of-speech tag; (vi) its classification as simple, complex or compound term; (vii) notes; (viii) one to five contexts for confirming whether the term is used in both languages in the same way; (ix) the term definition; (x) the IUPAC name for confirmation of the term.

The second part of our terminological record is completed with information about terms in Portuguese, as shown:



Table 4 – Part 2 of the terminological card.

Portuguese	
<b>Main equivalent:</b> malation (19,5)	<b>PoS:</b> substantivo
	<b>Type:</b> simples
<b>Variant I:</b> malationa (14,29)	<b>Notes:</b> Suffix -ona refers to the function called “cetona” (C=O) and the suffix -tiona refers to the function “tiona” (C=S). Interlinguistics causes
<b>Variant II:</b> malatião (3,57)	<b>Notas:</b> Portuguese from Portugal Dialectal and interlinguistics causes
<b>Contexts:</b>  Os compostos estudados têm em comum o grupo (P=S), mas após a adição de água de bromo, estes sofrem uma reação de oxidação com a formação de grupos oxons correspondentes (P=O), como no caso do clorpirifos e <b>malation</b> , que em geral potencializam a inibição da enzima acetilcolinesterase. <tesept00>  O inseticida <b>malationa</b> em calda oleosa é utilizado no controle do <i>Aedes aegypti</i> e a sua aplicação é feita por meio de nebulização. <artpt15>  O método emprega nanopartículas de ouro não modificadas, aptâmero específico para <b>malatião</b> e um péptido catiônico. <artpt64>	
<b>Definition in Portuguese:</b> Agrotóxico organofosforado do grupo dos fosforoditioatos.	
<b>IUPAC name in Portuguese:</b> 2-[(dimetoxifosforotioil)sulfanil]butanedioato de dietila	

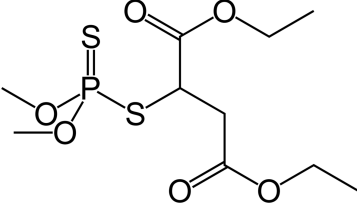
Source: the authors.

In fact, the second part of the record shows the same information as the first one, but now with the term in Portuguese taken from ORCHEUS corpus. As we can see in Table 4, we found three variants in Portuguese for the term malathion in English, which are “malation”, “malationa” and “malation”. The choice for the preferred equivalent has been following specific criteria, such as: a) to be the most frequent one in the corpus and the one that follows the guidelines of IUPAC; therefore, “malathion” was the final choice.

The last part of the record will show some important details for Chemistry students and specialists, as well as notes from the researcher who compiled the form. In Table 5, it is possible to see the last part of the record(i) the compound formula so researchers and students will be able to see its combinations either for confirmation or pedagogical purposes; (ii) notes about the term, for example, how it was found on the internet; (iii) the researcher’s name if the team needs to clear out any doubts about the form; (iv) the CAS number, so that consultants may

check additional information if necessary; (iv) notes about the variant the researcher considers important to be added; (v) the date when the term was described.

Table 5 – Part 3 of the terminological card.

Details	
 <p>Formula:</p>	CAS number: 121-75-5
Notes about the term: Existência da variante “malatiom” constatada na web.	Notes about the work: A variante “malatiom” não consta no corpus.
Researcher: José Victor de Souza	Date: 26/09/2022

Source: the authors.

## 5 Discussion

This study has produced two lists of 8,000 keywords each related to organophosphorus compounds. The group of researchers has produced two lists of seed words about organophosphorus terms, one in Portuguese and another one in Spanish, based on different methods in order to produce a new list of bilingual simple terms. The group has also led weekly meetings with the chemistry researcher and the terminologist to find similarities and differences described in the contexts found in the corpora in English and in Portuguese to better define the terms in this language.

Currently, we have been focusing on the finding of compounds specifically used as pesticides and creating a conceptual framework that categorizes them according to their chemical ligands. However, since organophosphorus compounds can have many different uses, our goal is to collect all OP compounds from our corpora, regardless of their usage as a pesticide or not, bearing in mind that some of those compounds may also have pharmaceutical properties. Besides that, other types of terms that are not necessarily compound names may be of our interest, such as adjectives and verbs. Then, we would finally be able to organize the most representative terms in a broader framework.

Regarding the second main question of this paper (2) How to prepare a bilingual descriptive digital glossary in a field with high denominational variation such as Pesticides Chemistry?, some aspects have called our attention concerning the terms in Portuguese. We can notice that the pattern of having two, three or more variants in Portuguese to the same term in English is still frequent. One example of that is the term dichlorvos, which is an organophosphate used as an insecticide to control household pests in public health, and protecting stored products from insects. In this case, we have found five variants in the ORCHEUS corpus in Portuguese: “diclorvós”, “diclorvos”, “DDVP”, “dichlorvos” and “diclórvos”. This result, as well as other organophosphorus pesticides have shown there is an insecurity by researchers who write their academic texts in Portuguese which is illustrated by the different accent for the same word or even no accent at all and even the option of using an acronym as well.

Concerning the concordance lines with the terms in context, another aspect that has called our attention is the fact that some pesticides are used in Brazil and prohibited in the U.S or European countries. This information has been added to the terminological forms by the research group to be discussed with the chemistry specialists to better define the terms in Portuguese.

As for the third question we have asked in this paper, (3) How can the data collected from our corpus contribute to the teaching, in English and Portuguese, of those terms? We have observed that, by reading several concordance lines from the corpus which discuss the terms selected in the main corpus, we find rich and useful information that could be used pedagogically with students of Chemistry. This experience has been tested by two of the main researchers of this study. One of them used concordance lines in Portuguese during a Portuguese as Academic course for students of Chemistry. In the same course, another researcher who has been working with English also discussed, with the same group, ways of finding the best equivalents in English for some terms in Portuguese. In this case, the students accessed ORCHEUS corpus in Portuguese to read about the chosen terms. By doing so, they reported how much they had learned about each compound they were searching for and mentioned that this discipline, which was being offered in their fourth year of undergraduate studies, would be very helpful if it were taught in their first year. This feedback from the

Chemistry students has shown how relevant it is to have them read concordance lines of texts in their own area that points out to a possibility of using a Data-driven learning approach (JOHNS, 1986; BOULTON, 2010; CROSTHWAITE, 2020) with those students. We have observed so far that students of Chemistry and students of Translation can develop their knowledge of Chemistry as well as their bilingual competence when dealing with contexts in both languages. They must use critical thinking before deciding which contexts they can use in their own texts, as well as the ones to be selected as part of our glossary. Therefore, in a pedagogical perspective, students will learn a lot by being exposed to short versions of terms in their area.

This study is still being developed but we will conclude the main aspects of it in the next section.

## 6 Conclusions

In this paper we presented a study of the bilingual terminology of Pesticide Chemistry since it is an important aspect related to the Brazilian Economy. Although much has been studied about pesticides worldwide, the registration of terms in Portuguese is still problematic, since most of the publication in this area has been produced in English. The International Union of Pure and Applied Chemistry (IUPAC) has published guidelines for researchers to follow when referring to terms of Chemistry, however, as we have observed in our study, when it comes to using those terms in Portuguese, different variants are found in academic texts, such as the terms malathion and dichlorvos, in English, that have more than three registrations in Portuguese, what has led to the lack of the standardization required by IUPAC. This fact may cause misinterpretation not only by Chemistry specialists, students and translators but also by general society, specially farmers who have bought and used these compounds in their crops and that could suffer from health harm if they use a pesticide that is not permitted by federal health agencies. With the aim of helping these consultants and users, we have been developing a terminological study in order to publish a bilingual glossary (English ↔ Portuguese) based on a combined methodology which involves corpus-based translation and terminology as well as natural language processing. In this case, an international partnership between São Paulo State University and the University of St. Gallen was set to develop an automatized methodology for retrieving terms in Portuguese based on a list of seed terms in English. The Brazilian and Swiss

group met in November, 2022, to run several trials to find a more refined methodology for term retrieval in different languages that can be used, in the future, by other areas.

Also, in order to produce a descriptive study, we selected simple terms in English and in Portuguese, generated from two corpora of academic texts in both languages, and showed them to a specialist in Chemistry, who has explained to the research group how and why those terms are used. These terms have been registered in terminological forms, together with their contexts of use, so that they can be part of the bilingual glossary with accurate definitions and information to help their consultants the best way possible.

Finally, we have concluded that reading concordance lines to be added to term definitions can be a pedagogical approach to students of Chemistry and Translation, since it combines critical criteria to understand which compound is being described as well as the best linguistic way to register this information in two different languages.

This research has shown that further investigations in the field of Pesticide Chemistry terminology are needed since it is a key area of Brazilian economy that has not been closely observed when it comes to language registration. We hope this work may encourage future studies to be carried out in this field.

### **Acknowledgements**

The authors would like to acknowledge the support by the Leading House (Univ. of St. Gallen), FAPESP (Process:19/14752-0; 21/08830-9) and Cnpq (Process:307287/2021-1).

### **References**

ANDRADE, M. M. de. Lexicologia, terminologia: definições, finalidades, conceitos operacionais. *In*: OLIVEIRA, A. M.; ISQUERDO, A. N. (org.). **As ciências do léxico: lexicologia, lexicografia, terminologia**. 2. ed. Campo Grande: Editora UFMS, 2001. p. 189-198.

AZENHA JR., J. **Tradução técnica e condicionantes culturais: primeiros passos para um estudo integrado**. Humanitas, 1999.

BARROS, L. A. **Curso básico de terminologia**. São Paulo: Edusp, 2004.

BOULTON, A. Data-driven learning: Taking the computer out of the equation. **Language learning**, v. 60, n. 3, p. 534-572, 2010.

BOWKER, L.; PEARSON, J. **Working with specialized language: a practical guide to using corpora.** Amsterdam: Routledge, 2002.

CABRÉ, M. T. **La terminología: Representación y comunicación: elementos para una teoría de base comunicativa y otros artículos.** Madri: Institut Universitari de Lingüística Aplicada (IULA), 1999a. <https://dialnet.unirioja.es/servlet/libro?codigo=64860>.

CABRÉ, M. T. **Terminology: theory, methods, and applications.** Amsterdam: John Benjamins Publishing, 1999b.

CABRÉ, M. T.; CONDAMINES, A.; IBEKWE-SANJUAN, F. Introduction: Application-driven terminology engineering. *In*: CABRÉ, M. T.; CONDAMINES, A.; IBEKWE-SANJUAN, F. (Org.). **Application-driven terminology engineering.** Amsterdam: John Benjamins, 2007. p. 1-17.

CAMARGO, D. C. D.; ROCHA, C. F.; PAIVA, P. T. P. **Pesquisas em estudos da tradução e corpora eletrônicos no Brasil.** São José do Rio Preto: Editora Unesp, 2012.

CONDAMINES, A. Variations in terminology: Application to the management of risks related to language use in the workplace. **Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication**, v. 16, n. 1, p. 30-50, 2010.

COSTA, M. D. M. **Avaliação da resistência a inseticidas e mecanismos selecionados em populações de *Aedes aegypti* Linnaeus 1762 (Diptera, Culicidae) da fronteira entre Brasil e Guiana Francesa.** 2017. Tese (Doutorado em Biologia Parasitária) – Fundação Oswaldo Cruz, Rio de Janeiro, 2017.

CROSTHWAITE, P. Taking DDL online: Designing, implementing and evaluating a SPOC on data-driven learning for tertiary L2 writing. **Australian Review of Applied Linguistics**, v. 43, n. 2, p. 169-195, 2020.

DEMKOWICZ, S.; RACHON, J.; DAŠKO, M.; KOZAK, W. Selected organophosphorus compounds with biological activity. Applications in medicine. **RSC Advances**, v. 6, n. 9, p. 7101-7112, 2016.

EMSLEY, J. **Nature's building blocks: an AZ guide to the elements.** Oxford: Oxford University Press, 2011.

ESTEVES, M. B. **Um estudo sobre a equivalência conceitual entre termos do português do Brasil e do inglês: aspectos lexicais e semânticos.** 2010. Dissertação de Mestrado em Linguística, Universidade de Brasília, Brasília, 2010.

GUBELMANN, R.; HANDSCHUCH, S.; NIKLAUS, C.; SOUZA, J. V.; LIMA, M. F.; SERPA, T.; PINTO, P. T. Cross-Lingual Retrieval of Organophosphorus Pesticide Names in Brazilian Research Articles. (forthcoming).

FAULSTICH, E. Aspectos de terminologia geral e terminologia variacionista. **Tradterm**, v. 7, p. 11-40, 2001. <https://doi.org/10.11606/issn.2317-9511.tradterm.2001.49140>.

FAULSTICH, E. Variação em terminologia: aspectos socioterminologia. *In*: FAULSTICH, E. **Panorama actual de la Terminología**. Granada: Comares, 2002. p. 65-92.

HALLIDAY, M. A. New ways of meaning: the challenge to applied linguistics. *In*: PÜTZ, M. (ed.). **Thirty years of linguistic evolution**. Amsterdam: John Benjamins, 1992. p. 59-95.

HALLIDAY, M. A. K. **An introduction to functional grammar**. 2. ed. London: Edward Arnold, 1994.

JOHNS, T. Micro-concord: A language learner's research tool. **System**, v. 14, n. 2, p. 151-162, 1986.

KILGARRIFF, A. Simple maths for keywords. *In*: MAHLBERG, M.; GONZÁLEZ DÍAZ, V.; SMITH, C. (ed.). **Proceedings of Corpus Linguistics Conference CL 2009**. Liverpool: University of Liverpool, 2009.

KILGARRIFF, A. *et al.* The Sketch Engine: ten years on. **Lexicography**, v. 1, n. 1, p. 7-36, 2014. <https://doi.org/10.1007/s40607-014-0009-9>.

KRIEGER, M. G.; FINATTO, M. J. B. **Introdução à Terminologia: teoria e prática**. São Paulo: Contexto, 2004. <https://books.google.com.br/books?id=qaMoExUpoDwC>.

KRIEGER, M. da G.; SANTIAGO, M. S. Estudos de Terminologia para a tradução técnica. **Revista de Letras**, v. 2, n. 33, art. 33, 2014.

LO, C.; SIMARD, M. Fully unsupervised crosslingual semantic textual similarity metric based on bert for identifying parallel data. *In*: **Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)**. Hong Kong: Association for Computational Linguistics, 2019. p. 206-215. <https://doi.org/10.18653/v1/K19-1020>.

MCNAUGHT, A. D. **Compendium of chemical terminology**. Oxford: Blackwell Science, 1997.

PAIVA, P. T. P.; CAMARGO, D. C. D.; XATARA, C. M. Uma reflexão sobre a elaboração de um léxico bilíngüe preliminar na subárea de cardiologia a partir do uso de termos encontrados em um corpus paralelo e em dois corpora comparáveis. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, v. 24, p. 1-22, 2008. <https://doi.org/10.1590/S0102-44502008000100001>.

PAIVA, P. T. P. **Uma investigação de traduções de textos da área médica sob a luz dos estudos da tradução baseados em corpus**. Ph.D. Thesis. São Paulo State University, 2009.

PINTO, P. T.; LIMA, M. de F. A tradução na área de química orgânica: Da adaptação à tradução literal. **Estudos Linguísticos (São Paulo. 1978)**, v. 47, n. 2, p. 573-585, 2018. <https://doi.org/10.21165/el.v47i2.2050>.

QASEMIZADEH, B.; HANDSCHUH, S. The ACL RD-TEC: a dataset for benchmarking terminology extraction and classification in computational linguistics. *In: Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*. Dublin: CompuTerm, 2014. p. 52-63.

ROCHA, C. F.; LIMA, M. F.; SERPA, T. Uma terminologia bilíngue para a química de compostos organofosforados: um estudo baseado no uso de corpora na composição de glossários de linguagem de especialidade em realidade aumentada. *In: FELIZARDO, A. B.; SILVA, E. B.; and FIGUEIRA-BORGES, G. (org.). Linguagem e Ensino em Percursos Interculturais*. 1. ed. Campinas: Pontes Editores, 2020.

SARDINHA, T. B. Lingüística de Corpus: Histórico e problemática. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, v. 16, p. 323-367, 2000. <https://doi.org/10.1590/S0102-44502000000200005>.

SERPA, T.; CERNA-CHÁVEZ, R. Una terminología trilingüe para la Medicina Regenerativa y la Ingeniería de Tejidos: un estudio basado en el uso de corpora en la composición de glosarios de lenguaje especializado en realidad aumentada. **Revista Digital Internacional de Lexicología, Lexicografía y Terminología**, n. 3, p. 1-21, 2020.

SERPA, T.; CERNA-CHÁVEZ, R. Una terminología trilingüe para la Medicina Regenerativa y la Ingeniería de Tejidos: un estudio basado en el uso de corpora en la composición de glosarios de lenguaje especializado en realidad aumentada. **Revista Digital Internacional de Lexicología, Lexicografía y Terminología**, n. 3, p. 161-181, 2020.

SHARPLEY, A.; JARVIE, H.; FLATEN, D.; KLEINMAN, P. Celebrating the 350th anniversary of phosphorus discovery: A conundrum of deficiency and excess. **Journal of environmental quality**, v. 47, n. 4, p. 774-777, 2018.

SINCLAIR, J. **Corpus, concordance, collocation**. Oxford: Oxford University Press, 1991.

SOUZA, J. V. de. A questão terminológica dos organofosforados na química de pesticidas: Uma abordagem baseada em corpus. **Estudos Linguísticos (São Paulo. 1978)**, v. 48, n. 3, p. 1620–1638, 2019. <https://doi.org/10.21165/el.v48i3.2270>.

SOUZA, J. V. de; PINTO, P. T.; LIMA, M. M. de F. Malationa, malation ou malatiom? A variação denominativa no processo de criação de um glossário bilíngue da área de química de pesticidas. **Acta Scientiarum. Language and Culture**, v. 44, n. 1, e55894, 2022. <https://doi.org/10.4025/actascilangcult.v44i1.55894>.



SOUZA, J. V. de. **Uma proposta de vocabulário bilíngue de pesticidas organofosforados por meio da linguística de corpus**: foco no trato da variação denominativa. Master's dissertation (forthcoming). Universidade Estadual Paulista, São José do Rio Preto, 2023.

TOGNINI-BONELLI, E. **Corpus linguistics at work**. Amsterdam: John Benjamins, 2001. <https://benjamins.com/catalog/scl.6>.

Received on: 09.03.2023

Approved on: 07.04.2023