# Corpus-Based Methods and AI-Assisted Terminography for Contextonym Analysis

**Antonio San Martín**

University of Quebec in Trois-Rivières, 3351, boulevard des Forges,
Trois-Rivières (Quebec) G8Z 4M3 Canada
E-mail: antonio.san.martin.pizarro@uqtr.ca

**Abstract**

This paper presents contextonym analysis as a hybrid method combining corpus-based techniques and generative artificial (GenAI) tools to support the writing of precise, context-sensitive terminological definitions. Grounded in the Flexible Terminological Definition Approach, this method is based on the premise that definitions should reflect the most relevant conceptual content activated in specific contexts. Contextonyms (frequent surface co-occurrents within a 50-word window) are extracted in word sketch (WS) form in Sketch Engine and help reveal salient semantic features of a target term without relying on predefined syntactic or semantic relations. The paper outlines strategies for interpreting contextonyms, including filtering concordance lines, consulting WSs, and prompting GenAI tools to assist with interpretation. A typology of contextonyms is proposed, along with a case study illustrating how the method supports the creation of domain-specific definitions. By combining corpus data with AI-assisted interpretation, contextonym analysis offers a robust and user-friendly approach to terminological definition writing.

**Keywords:** contextonym; terminological definition; word sketch; AI-assisted terminography

## 1. Introduction

Definitions are a key component of terminological resources. Consequently, terminologists need user-friendly corpus-based techniques to identify the most relevant content for terminological definitions. To this end, San Martín (2016) introduced contextonym analysis. Inspired by distributional semantics (Lenci, 2008), it involves analyzing the terms that frequently co-occur with the target term, regardless of their syntactic relationship. These co-occurring terms, called contextonyms[1], help identify the most relevant semantic features of a term[2].

[1] The term *contexonym* was coined by Ji et al. (2003). We prefer the variant *contextonym*, used by other authors (Gadek et al., 2017; Șerban et al., 2012), as it better highlights the link to the notion of context.

[2] While this paper focuses on contextonym analysis for writing terminological definitions, this technique can also be valuable for identifying conceptual relations for terminological knowledge bases and ontologies.

For example, the most frequent contextonyms of *nematode* in an agricultural corpus[3] are *plant, soil, crop, root, disease, population, pest*, and *yield*. Examining these contextonyms and associated concordance lines reveals key semantic features: some nematode species act as pests by damaging plant roots and spreading soil-borne diseases, reducing crop yield, while other nematode species help regulate pest populations and support soil health.

An empirical study (San Martín, 2025) found that the optimal window size for extracting contextonyms to define terms is 50 words beyond sentence boundaries. In this context, a contextonym is defined as a word that frequently appears within 50 words of the target term. To offer a user-friendly extraction method, the study provided a custom sketch grammar for obtaining contextonyms in the form of a word sketch (WS) column in Sketch Engine (SkE).

While contextonym extraction can now be easily performed with SkE, guidelines on its application by terminologists remain lacking. To address this gap, this paper explores how to analyze a term's contextonyms to identify the most relevant semantic features for inclusion in a terminological definition.

The paper is structured as follows: Section 2 introduces the Flexible Terminological Definition Approach, which forms the basis of this study, along with various methods for definition writing. Section 3 explains how contextonyms can be extracted using WSs, with particular attention to the study that established the optimal extraction parameters. Section 4 discusses the interpretation of contextonyms, the different types that may be identified, and the advantages and limitations of the approach. Section 5 presents a case study on contextonym-based definition writing. Finally, Section 6 offers concluding remarks.

## 2. Terminological Definition Writing

### 2.1 The Flexible Terminological Definition Approach

The classical approach[4] to terminological definitions, known as the intensional or analytic definition, is based on the specification of the necessary and sufficient features of the concept denoted by the term to be defined. This approach assumes that such features are universal and context-independent. However, as Cognitive Linguistics has shown, it is often impossible to objectively determine these features because concepts are fuzzy and lack clear boundaries (Temmerman, 2000: 7). Moreover, intensional

[3] Unless otherwise indicated, all examples come from an 8-million-word corpus of English agricultural texts. For details regarding its composition, see San Martín (2025, p. 7).

[4] This approach remains widely advocated by terminology manuals (Dubuc, 2002; Kockaert & Steurs, 2015; Suonuuti, 1997), the ISO 704:2022 standard on terminology work (ISO/TC 37/SC 1, 2022), and specialized manuals on terminological definitions (Fargas, 2009; Vézina et al., 2009).

definitions are often less helpful for non-experts, as they omit encyclopedic knowledge that aids concept understanding.

To address these limitations, the Flexible Terminological Definition Approach (FTDA) (San Martín, 2016, 2022a) proposes a shift in how definitions are constructed, as an alternative to the classical approach. Instead of focusing on essentialist criteria, the FTDA adopts a usage-based, context-sensitive view of meaning. In line with scholars such as Temmerman (2000: 43) and Seppälä (2015: 33), the FTDA replaces the idea of necessary and sufficient features with that of relevant features (i.e., features that emerge as salient in a given context of use). Consequently, a terminological definition is a natural-language description of the most relevant conceptual content conveyed by a term.

According to the FTDA, crafting definitions that meet user needs requires considering the role context plays in meaning construction. From a cognitive linguistics standpoint, terms do not possess meaning but function as access points to large networks of knowledge (Evans, 2019: 392). It is context, understood as any factor influencing interpretation (Kecskes, 2023: 26), that determines which segment of this knowledge (i.e., which meaning) is activated in each usage event.

All the knowledge that a term is capable of activating is its semantic potential (Evans, 2015; Hanks, 2020). Semantic potential includes the associated concept (or concepts, in the case of polysemy), along with all relevant frames (i.e., encyclopedic knowledge structures that organize and relate concepts within a particular scene, situation, or event (Evans, 2019)). For example, the semantic potential of *carbon*, a chemical element, encompasses all the knowledge that it can activate in any context. In contrast, meaning is the specific knowledge conveyed in each usage event (a narrow portion of the semantic potential), as in a tweet posted on 24 September 2024 by European Commission President Ursula von der Leyen[5], where *carbon* is conceptualized as a pollutant whose release must be financially compensated.

A definition cannot describe a term's semantic potential because it is too vast. Nor can it explain meanings because meanings are inherently transient and linked to specific usage events. When terminologists craft definitions, they must select the most relevant information from a term's huge semantic potential, narrowing it based on contextual constraints, which can be linguistic, thematic, cultural, ideological, geographical, and chronological (San Martín, 2022b). Applying contextual constraints results in a specific conceptual subset known as premeaning (Croft & Cruse, 2004: 110), which is what a terminological definition describes.

Examples of premeanings based on a thematic constraint are that of *ammonia* in Aquatic Ecology, where its role in oxygen depletion and toxicity to aquatic organisms

[5] "The European experience shows it: carbon pricing works. We are encouraging more countries to join the movement. And bring industry on board ↓"
(https://x.com/vonderleyen/status/1838650409906274452).

is emphasized, and its premeaning in the domain of Agriculture, where ammonia is valued as a nitrogen source in fertilizers. Premeanings can also be shaped by various contextual constraints simultaneously. For instance, in Hydraulic Engineering within a Dutch context, the premeaning of *flood* is linked to sea-level rise and river overflow, emphasizing advanced water management infrastructure such as dike systems, storm surge barriers, and pumping stations. In contrast, in Human Geography in Bangladesh, *flood* is primarily associated with seasonal monsoons and river basin overflow, shaping settlement patterns, migration, and socio-economic vulnerability in low-lying areas.

For definition purposes, a premeaning corresponds to a portion of a single concept and the corresponding frames. When a term is polysemic (i.e., linked to multiple concepts) it is customary to provide at least one definition for each distinct concept. Nonetheless, the FTDA contends that the provision of multiple definitions should not be limited to polysemic terms alone. Given the plurality of possible premeanings, a monosemic term may also have more than one definition in a single terminological resource so as to reflect contextual variation[6]. Accordingly, a single concept may be defined in different ways depending on the contextual constraints (or their combination) relevant to the specific terminological resource. For instance, in the case of *flood*, there could be several definitions from the different thematic and geographical perspectives within the same resource.

Several methods can be used to identify the information necessary to accurately describe a premeaning within a definition. However, to effectively determine how a term is conceptualized under specific contextual constraints, corpus analysis is essential. Importantly, the selected corpus must align with the contextual constraints imposed on the definition. As demonstrated below, contextonym analysis is an effective corpus method enabling the extraction of the most relevant semantic features of a term depending on particular contextual constraints.

### 2.2 Methods for Terminological Definition Writing

Terminologists can draw on both non-corpus-based and corpus-based methods to craft definitions. Traditional non-corpus approaches include consulting existing definitions, reviewing specialized literature, and seeking input from domain experts. More recently, generative artificial intelligence (GenAI) has emerged as a complementary tool in what can be referred to as *AI-assisted terminography* (San Martín, 2024: 4). GenAI can

[6] Contextual variation arises when a term does not consistently activate the same semantic features, whose salience varies depending on context. Also called conceptual variation (Freixa & Fernández-Silva, 2017), it is closely linked to polysemy. Although the two are not equivalent, their boundaries are often blurred in practice. Contextual variation entails the differential activation of features of a single concept across diverse contexts, whereas polysemy involves the activation of distinct concepts by the same term in different instances of use.

support definition writing in various ways, including post-editing terminography[7] (where the machine generates a draft definition that the terminologist refines): it can resolve notional doubts, analyze existing definitions, and evaluate and enhance definitions produced by the terminologist (San Martín, 2024: 4–6).

These approaches can be complemented by corpus-based methods, which allow for the analysis of large volumes of specialized texts simultaneously. Corpus analysis enables terminologists to observe how terms are actually used in real contexts. The most basic method for analyzing a corpus to define a term involves examining the concordance lines where the term occurs. However, this process can be highly time-consuming. A more efficient strategy consists in leveraging co-occurrence data to filter and prioritize concordance lines. The two primary types of co-occurrence employed to extract semantic data about terms are syntactic co-occurrence (i.e., collocations) and semantic co-occurrence (i.e., knowledge patterns). Contextonym analysis introduces a third co-occurrence type into the terminologist's toolkit: surface co-occurrence.

Syntactic co-occurrence occurs when two words share a direct or indirect syntactic relation in a given linguistic context (Evert, 2009: 1215). For instance, a noun and the verb of which it is the subject, such as *doctor* and *diagnose* in "Doctors diagnosed her with lupus". When a syntactic co-occurrence is recurrent, it is usually referred to as a *collocation*. Currently, collocational data can be readily extracted from any corpus with the WS function in SkE (Kilgarriff et al., 2014). A WS provides an overview of the most common usage patterns of a given search word in a corpus. Its various columns display the words that are syntactically related to the search word, along with their frequency, the association score logDice (Rychlý, 2008), and links to the corresponding concordances. Figure 1 provides examples of some WS columns offered by SkE by

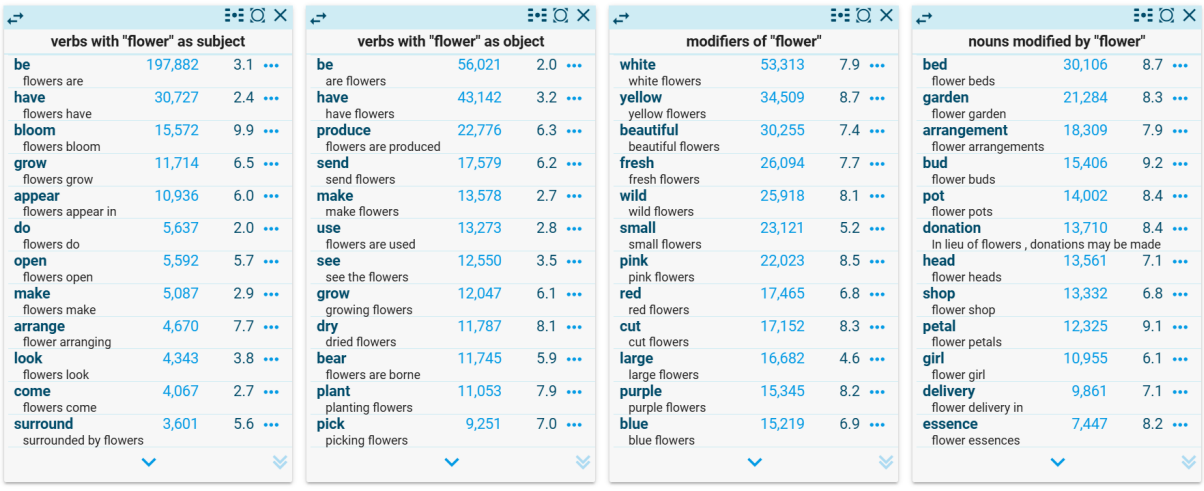


Figure 1: Default WS columns of the word *garden* in the English Web 2021 (enTenTen21) corpus

[7] The term *post-editing terminography* is an adaption of *post-editing lexicography* (Jakubíček et al., 2018).

default. While collocation analysis offers insights into the semantic behavior of terms, it is insufficient on its own for definition purposes. Syntactic co-occurrence needs to be complemented with other types of co-occurrence.

Semantic co-occurrence arises when two words share a semantic relation (e.g., hyponymy, meronymy, cause) within a given linguistic context (San Martín & Trekker, 2021, p. 66). Semantic co-occurrents can be identified in corpora using lexico-syntactic patterns conveying a specific semantic relation, known as knowledge patterns (Meyer, 2001). For example, the pattern "x *such as* y" (e.g., "fertilizers such as compost") conveys a hyponymic relation between x (*fertilizer*) and y (*compost*). Compared to syntactic co-occurrence, the conceptual information derived from semantic co-occurrence is more useful for definition writing. However, extracting meaningful results with knowledge patterns requires large corpora, and the process can be affected by significant noise (Bowker, 2003). Semantic co-occurrence can be extracted with the EcoLexicon Semantic Sketch Grammar (http://ecolexicon.ugr.es/essg) (León-Araúz et al., 2016; León-Araúz & San Martín, 2018), which generates WS columns for hyponymy,

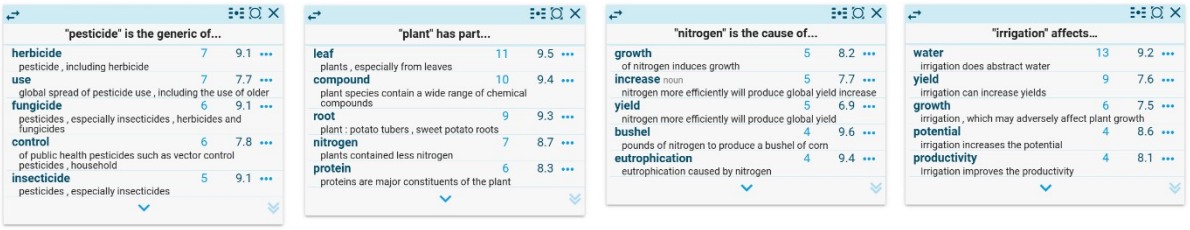


Figure 2: Semantic WS columns from the agricultural corpus

meronymy, cause, function, and location from any English user-owned corpus. A subsequent development added a WS column for the agent-patient relation (*affects/is affected by*) (San Martín et al., 2023). Figure 2 provides examples of semantic WS columns.

To overcome the limitations of syntactic and semantic co-occurrence, San Martín (2025) proposes surface co-occurrence for extracting semantic data to support terminological definition writing. Surface co-occurrence occurs when two words appear within the same linguistic context, regardless of whether a syntactic or semantic relationship exists between them (Evert, 2009, p. 1215). For example, in "…potash application on soils low in magnesium…", the terms *potash* and *magnesium* are surface co-occurrents. Analogous to how frequent syntactic co-occurrents are referred to as collocations, the term *contextonym* denotes frequent surface co-occurrents (San Martín, 2025, p. 5). The main limit for considering that two terms are contextonyms is the window size or number of words between them. Other constraints may include restricting analysis to specific word classes (typically nouns, adjectives, and verbs) and applying a stoplist to exclude certain words.

Contextonym analysis can yield relevant insights even from smaller corpora than semantic co-occurrence methods. It can also uncover a wider range of semantic

information without the need to manually define the relations to capture.

As explained in the next section, San Martín (2025) developed a sketch grammar to extract contextonyms with the WS function in SkE.

## 3. Contextonym Extraction with Word Sketches

Since the contextonyms of a term are meaningful indicators of its semantic value (Ji et al., 2003: 194), contextonym extraction represents a valuable technique for supporting terminological definition writing. Contextonymy is not a transitive relation (e.g., the fact that *irrigation* is a contextonym of *plant* and that *furrow* is a contextonym of *irrigation* does not imply that *furrow* is a contextonym of *plant*), nor is it a symmetric relation (e.g., the fact that *pesticide* is a contextonym of *pollinator* does not necessarily imply that *pollinator* is a contextonym of *pesticide*). Furthermore, contextonyms, unlike hypernyms or synonyms, do not necessarily belong to the same word class (e.g., the adjective *green* can be a contextonym of the noun *manure*) (Ji et al., 2003: 195).

The contextonyms of a term can reflect various semantic relations, both hierarchical (i.e., hypernymy, meronymy) and non-hierarchical (e.g., cause, function, location), as well as domain-specific relations (e.g., *is a pest of* in Agriculture). Contextonyms can also be participants in the same frame. For example, the term *pesticide* activates the frame of *pest management*, and many of its contextonyms (such as *pest*, *crop*, *farmer*, *risk*, *health*, and *control)* are elements within that frame.

Contextonyms differ depending on the contextual parameters at play. For example, the contextonyms of *chlorine* in an Air Quality Management corpus (*ozone*, *stratosphere*, *CFC*, *depletion*, etc.) describe it as a contributor to stratospheric ozone depletion, whereas its contextonyms in a Water Treatment corpus (*water*, *disinfection*, *chlorination*, *kill*, etc.) emphasize its function as a water disinfectant (San Martín, 2022a: 68). Consequently, the corpus used must align with the contextual constraints of the definitions to be created.

The parameters used to extract contextonyms can vary. Consequently, San Martín (2025) conducted an experiment to determine the optimal configuration for WS-based contextonym extraction in terminological definition writing. The parameters considered included the window size, whether sentence boundaries should be crossed, and how the results should be ordered (by frequency or by association score). The study adopted as default parameter the restriction that contextonyms be limited to nouns, adjectives, and verbs. In addition, common words deemed too semantically general to aid in definition writing (e.g., *be*, *do*, *other*, *same*) were excluded.

For the experiment, 20 agricultural terms in English were selected. For each term, definitions were gathered from various sources to create 20 corpora of definitions. The most frequent terms in each definition corpus served as a gold standard. Using an

agricultural corpus, the contextonyms of each term were extracted based on different window sizes. These contextonyms were then compared (using cosine similarity and precision) with the list of the most frequent terms extracted from the corresponding definitions. Cosine similarity was chosen for its ability to account for the ordering of contextonyms, while precision focused on their presence or absence. The final results were based on a combination of both metrics.

The experiment found that the optimal window size for extracting contextonyms in the form of a WS column to support terminological definition writing is 50 tokens, allowing sentence boundaries to be crossed. The study also concluded that contextonyms should be ranked by frequency rather than by association score.

The custom sketch grammar designed to generate the WS column for contextonym extraction is available in San Martín (2025: 17) along with instructions on how to use it with any user-owned corpora in SkE.

## 4. Contextonym Analysis

### 4.1 Contextonym Interpretation

By default, contextonyms in a WS column are ordered by association score, each accompanied by a short textual fragment (called "collocation example" or "longest-commonest match") showing the most frequent way the target term and the contextonym co-occur within a short distance. As previously mentioned, users should order the results by frequency (Figure 3), as the association score has proven unhelpful for definitional purposes (San Martín, 2025: 14). The ordering of contextonyms is informative, as the most frequent contextonyms are potentially more relevant for defining the term.

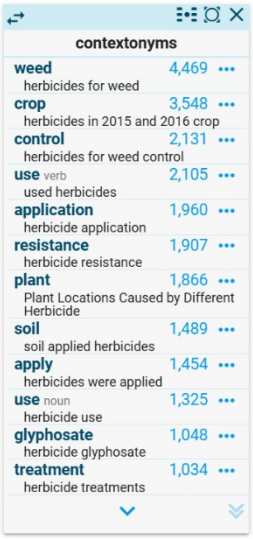


Figure 3. Contextonym WS column of the term *herbicide*

Since contextonym extraction is not based on predefined syntactic or semantic relations, terminologists must interpret the nature of the relationship linking each contextonym to the target term. To assist with this, several strategies can be employed, falling into two main categories: corpus-based methods and GenAI methods.

### 4.1.1 Corpus-based methods

The most straightforward way to interpret contextonyms is by examining the corresponding concordance lines. In SkE, users can easily access these lines for any contextonym. However, contextonyms are often linked to numerous concordance lines, making it necessary to apply a method for selecting the most informative ones.

Before consulting the concordance lines associated with a given contextonym, users can refer to the longest-commonest match displayed beneath the contextonym in the WS column. This short excerpt may offer insight into the relationship between the target term and the contextonym. For example, in Figure 4, *soil* appears as a contextonym of *rhizobium*, and the longest-commonest match indicates that rhizobia can be found in soil. However, many longest-commonest matches are uninformative (e.g., the longest-commonest match of *plant* in Figure 4), making the consultation of concordance lines necessary.

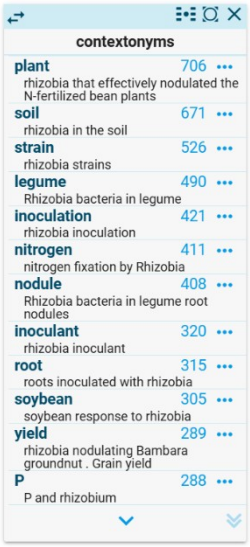


Figure 4: Contextonym WS column of the term *rhizobium*

With San Martín's (2025) sketch grammar, contextonyms can appear up to 50 words away from the target term, even across sentence boundaries. As a result, in many concordance lines, the contextonym and target term may appear too far apart to show a clear relationship. To overcome this difficulty, users can retain only concordance lines where both terms appear within the same sentence. Users can further filter lines by keeping only those where both terms are at most five words apart, increasing the

likelihood that the lines will clearly reflect the relationship between the two terms. Finally, repeated lines[8] can be removed with the "Hide sub-hits" option.

Nonetheless, even after applying these filters, a substantial number of concordance lines may remain. In this context, the GDEX (Good Dictionary Examples) function in SkE (Kilgarriff et al., 2008) is particularly useful. GDEX ranks lines based on a score derived from various criteria determining their suitability as dictionary examples. These criteria include the requirement that lines be complete sentences, a preference for short sentences with common vocabulary, and penalties for sentences containing pronouns, anaphoric elements, or words with non-alphabetic symbols. Though not designed for contextonym interpretation, GDEX effectively aids in identifying lines where the relationship between the contextonym and the target term is likely to be clear.

For example, *weed* is a contextonym of *glyphosate*. As shown in Table 1, the filtered concordance lines ranked by GDEX score allow us to infer that glyphosate is used to control weed growth, although weed resistance to glyphosate is increasing.

| **Sentence** | **GDEX score** |
|---|---|
| **Glyphosate** or glufosinate ammonium restricted **weed** mass more than the alternative treatments except flaming or mulching. | 0.95 |
| The development of **glyphosate** resistance in **weed** species is emerging as a costly problem. | 0.95 |
| **Glyphosate** gave strong suppression of **weed** growth at the end of the season. | 0.949 |
| Saflufenacil was introduced in part to manage **glyphosate** - resistant **weeds**. | 0.925 |
| In some areas **glyphosate** resistant **weeds** have developed, causing farmers to switch to other herbicides. | 0.925 |

Table 1: Concordance lines ranked by GDEX score

Another corpus-based approach to interpret contextonyms consists of verifying if they also appear in other WS columns (both default WS columns and the ones generated with the EcoLexicon Semantic Sketch Grammar). For instance, *herbicide* is a contextonym of *atrazine* and can also be found in the *"atrazine" is a type of...* WS column, confirming it as the hypernym of *atrazine*.

For verb contextonyms, other WS columns help determine whether the target term functions as a subject or object of the contextonym or holds another kind of relationship with the contextonym. For instance, *apply* is a contextonym of *herbicide*, and its presence in the WS column *verbs with "herbicide" as object* confirms that relationship. In contrast, although *grow* is a contextonym of *germination*, its absence in other WS columns of *germination* indicates that the relationship is neither subject nor object. In

[8] Repeated lines can happen because around a given target term, the same contextonym can appear more than once.

fact, associated concordance lines reveal that germination is a process occurring before plants grow.

For adjective contextonyms, other WS columns help determine whether the adjective typically modifies the target term or holds a different type of relationship. For example, *organic* is a contextonym of *fertilizer*, and its appearance in the *modifiers of "fertilizer"* column confirms a modifier relationship. In contrast, while *rhizomatous* is a contextonym of *glufosinate*, it does not appear in any other WS columns for *glufosinate*, suggesting that it does not modify it directly. Concordance lines further clarify that *rhizomatous* modifies *johnsongrass*, and the connection to *glufosinate* lies in the fact that glufosinate is an herbicide used to control rhizomatous johnsongrass.

However, while consulting other WS columns can sometimes reveal the relationship between a target term and its contextonyms, this method is not always efficient. Some contextonyms appear in multiple WS columns and reviewing the corresponding concordance lines to identify which WS best reflects the relationship can be time-consuming. Moreover, WS columns may include noise, further reducing the method's efficiency. For example, *herbicide* is a contextonym of *weed* and appears in 11 syntactic WS columns (which includes 8 prepositional ones such as *... to "weed"* or *"weed" with ...*) and 8 semantic WS columns. Some of these are incorrect due to noise.

Regardless of the method used, one key challenge is that the semantic link between a target term and its contextonyms is often multifaceted and complex. It frequently does not align with standard semantic relations such as hyponymy or cause. For example, *sulfur* is a contextonym of *potash*, but their relation is too intricate to be captured in a single sentence[9]. This complexity also makes analyzing concordance lines more time-consuming.

As shown below, GenAI can help streamline relation identification between a target term and its contextonyms.

### 4.1.2 AI-Assisted Terminography

GenAI's potential for terminology work has led to AI-assisted terminography (San Martín, 2024, p. 4), which includes all forms of support that AI tools can offer to terminologists. In contextonym analysis, GenAI tools such as ChatGPT or Gemini can assist in two main ways. The first involves prompting the GenAI tool to explain the relationship between a target term and its contextonyms, based on the knowledge embedded in its training data. Although this method is quick and efficient, it can produce inaccurate information (i.e., hallucinations), especially when the language

[9] Potash contains potassium, an essential nutrient for plant growth. Sulfur is another vital nutrient for plants. Potassium and sulfur often work synergistically in plant metabolism, as sufficient potassium improves sulfur uptake and vice versa. Fertilizers like sulfate of potash provide both nutrients, supporting balanced plant nutrition.

model lacks sufficient knowledge about the subject[10]. Table 2 provides an example of a prompt that can be used to obtain an explanation of the relationship between a target term and its contextonyms.

To potentially reduce hallucinations, GenAI tools can be prompted to base responses on web searches. However, this method has limitations, as it typically cannot provide a source for each individual term due to constraints on the number of web results retrieved. To address this, terminologists can use the Deep Research function available in some GenAI tools, which draws on dozens of web sources to generate a response. In our preliminary tests with ChatGPT, Gemini, and Grok, we were only able to obtain a list of explanations with a source attributed to each contextonym with ChatGPT. Furthermore, ChatGPT 5 thinking model, which combines reasoning capabilities with web searches, is also able to generate explanations for each contextonym, each linked to a corresponding source. The prompt used is reproduced in Table 3.

| |
|---|
| The following is a list of terms related to "rhizosphere". For each term, provide a one- or two-sentence explanation describing its relationship to "rhizosphere" in an agricultural context.<br><br>*[List of contextonyms]* |

Table 2: Prompt to obtain an explanation of the relation between a target term and its contextonyms

| |
|---|
| The following is a list of terms related to "glyphosate". For each term, provide a one- or two-sentence explanation describing its relationship to "glyphosate" in an agricultural context. For each explanation, provide the source.<br><br>*[List of contextonyms]* |

Table 3: Prompt to obtain an explanation of the relation between a target term and its contextonyms with Deep Research or ChatGPT 5 thinking model

The other way in which GenAI tools can assist terminologists is by analyzing concordance lines to identify the relationship between contextonyms and the target term. For example, *nitrogen* is a contextonym of *biomass,* with 476 associated lines in our agricultural corpus. After applying the previously described filters (i.e. keeping only those where both terms are in the same sentence and at most five words apart, and removing repeated lines), 35 lines remain. These can be submitted to the GenAI tool to generate an explanation of the relationship between *nitrogen* and *biomass* (Table 4). This approach reduces the risk of hallucination. The main drawback is that copying and pasting the lines for each contextonym is more time-consuming than prompting the GenAI tool to explain the relationship based solely on its training data.

[10] It is important to note that no hallucinations were detected when using the example prompts in this paper. However, since the terms used are central to the domain of Agriculture, they are undoubtedly well represented in ChatGPT's training data.

Nonetheless, it is still faster than manual analysis.

| |
|---|
| The following is a set of concordance lines containing the terms “nitrogen” and “biomass”. Analyze these lines to explain the relationship between the two terms in an agricultural context.<br><br>*[Concordance lines]* |

Table 4: Prompt to obtain an explanation of the relation between a target term and one of its contextonyms based on a set of concordance lines

Given the specific advantages and limitations of both corpus-based and GenAI-based approaches, a combined strategy is the most effective. GenAI tools provide a rapid, albeit potentially unreliable, overview of the relationship between a target term and its contextonyms. This output can serve as a starting point for terminologists, who can then validate, refine, or expand using corpus-based methods.

### 4.2 Contextonym Typology

If the target term is frequent enough in the corpus, most of its contextonyms will reflect its relevant features, and their co-occurrence frequency will indicate their level of relevance. In many cases, contextonyms also point to the conceptual frames in which the target term participates, offering insight into the broader knowledge structures activated by the term. However, terminologists must assess whether the information conveyed by each contextonym is relevant to the specific definition, as some may not be useful depending on the context or purpose of the resource in which the definition will be inserted.

One case where not all the information conveyed by contextonyms is included in the definition is in terminological resources where feature inheritance is applied to definitions. For instance, since *glyphosate* is a type of *herbicide*, its definition will use *herbicide* as genus. *Glyphosate* is then assumed to inherit the features associated with *herbicide*. As a result, the information present in the definition of *herbicide* is not repeated in the definition of *glyphosate*. In fact, terms linked by hyponymy tend to share contextonyms. For instance, out of the 30 most frequent contextonyms, *glyphosate* and *herbicide* share 19. However, even when feature inheritance is applied to definitions, it may still be appropriate to repeat certain information in both the hypernym and hyponym definitions to ensure user comprehension.

Among the contextonyms of a term, some may be too general to clearly convey a specific semantic relation (e.g., *include*, *use, provide*[11]). Many of these terms belong to

[11] In San Martín’s (2025) sketch grammar for extracting contextonyms, a list of words is excluded due to their irrelevance for definitional purposes (e.g., *be*, *do*, *same*, *other*). This stoplist follows a conservative approach, but users can adjust the sketch grammar to exclude additional terms.

the Transdisciplinary Scientific Lexicon (TSL)[12] (Drouin, 2010), a set of general-language words central to scientific discourse (e.g., *model*, *analysis*, *system)*. However, TSL units are not inherently irrelevant as contextonyms. In some cases, they highlight key semantic information of a term. For instance, *control,* a TSL unit, is a contextonym of *paraquat* and provides valuable information, as paraquat is an herbicide used to control weed species.

Finally, the results may also include noise, depending on the nature of the corpus. For example, in our agricultural corpus, *table* and *figure* appear as contextonyms of *genotype*. This is because the documents included in the corpus contain a considerable number of tables and figures, which are referred to in the texts.

The same contextonym may be relevant for the definition of one target term but not for another. Ultimately, through careful analysis, the terminologist must determine which contextonyms convey relevant information to include in a given definition. This decision depends not only on the nature of each contextonym but also on the intended user and the specific characteristics of the resource in which the definition will appear.

### 4.3 Advantages and Limitations of Contextonym Analysis with Word Sketches

A key advantage of contextonym analysis is its ability to show how a term is conceptualized within a specific context by using a corpus representative of that context. This enables the creation of definitions adapted to different contexts. For instance, to define *genotype* from the point of view of agriculture, we can analyze its contextonyms in a corpus of agricultural texts.

Despite GenAI's ability to generate definitions (San Martín, 2024), contextonym analysis offers a distinct advantage: it is grounded in verifiable, empirical data. Unlike GenAI, which generates content based on probabilistic language modeling that cannot be normally traced to specific sources, contextonym analysis relies on co-occurrence patterns extracted from a corpus. This means that every contextonym can be traced back to actual usage examples, allowing terminologists to consult concordance lines to assess relevance.

However, contextonym analysis in the form of WS columns also presents certain limitations. Although contextonyms can be extracted for multiword terms (albeit with some constraints), WS columns list contextonyms as individual words. For example, *Palmer* and *amaranth* both appear as contextonyms of *glyphosate*. However, it would be more accurate to identify the multiword term *Palmer amaranth* (a specific weed

[12] The list of TSL units can be consulted in <https://olst.ling.umontreal.ca/lexitrans/>.

species known for its resistance to glyphosate) as the contextonym. The lack of multiword grouping can make contextonym interpretation more time-consuming.

Another major limitation of contextonym analysis in the form of WS columns is that, in the case of polysemous terms, the contextonyms for all senses are combined. SkE addresses this through its "Show senses" function, which performs word sense induction by grouping WS results into different, automatically detected senses. Each induced sense is labeled with a set of representative words. While this functionality can be helpful, its effectiveness depends on two conditions: the senses must be distinct enough to allow for clustering, and frequent enough in the corpus to be statistically recognized.

## 5. From Contextonyms to Definitions: a Case Study

To illustrate the application of contextonym analysis in terminological definition writing, this section presents a case study on the term *methane*. Drawing on two specialized corpora in Waste Management and Energy Engineering, with approximately 1.2 million words each, the contextonyms of *methane* in each domain are extracted and interpreted to craft two context-sensitive definitions.

### 5.1 Definition of *methane* in Waste Management

Using the Waste Management corpus, the most frequent contextonyms of *methane* include *landfill*, *gas*, *waste*, $CO_2$, and *carbon* (Figure 5)

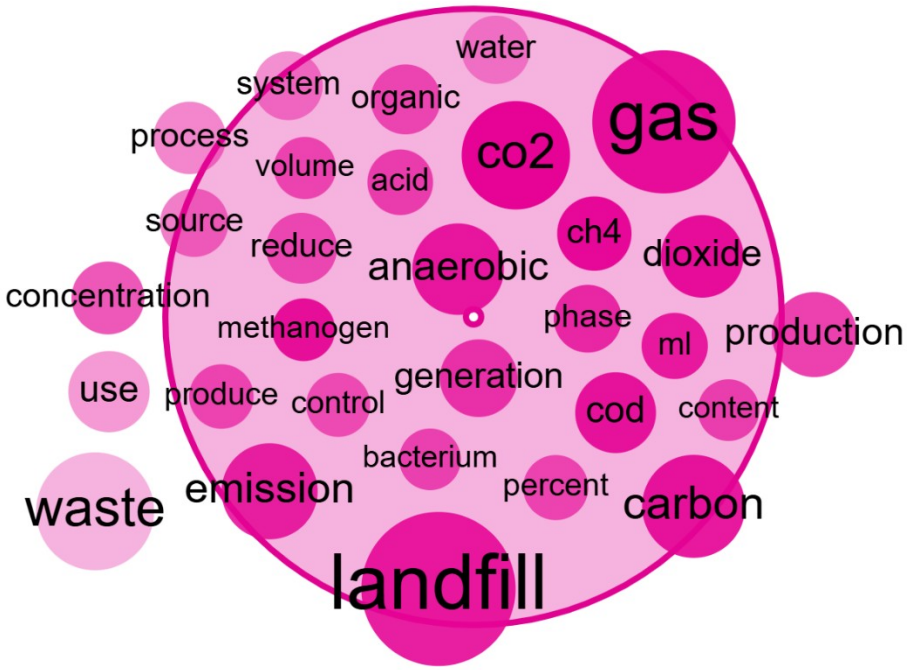


Figure 5: Visualization generated by SkE showing the 30 most frequent contextonyms of *methane* in the Waste Management corpus (bubble size indicates frequency)

These contextonyms show that *methane* is mainly conceptualized as a gas produced by anaerobic decomposition of organic waste, especially in landfills. Contextonyms such as *emission*, *$CO_2$*, *carbon*, and *dioxide* emphasize its environmental impact as a greenhouse gas contributing to climate change. Meanwhile, *methanogen*, *anaerobic*, *process*, and *acid* suggest the microbial and chemical pathways through which methane is produced in oxygen-free conditions. Contextonyms like *generation*, *production*, *system*, and *control* point to engineered processes for reducing methane emissions or capturing it as an energy source.

Based on the analysis of these contextonyms through the methods explained in the previous sections, a context-sensitive definition of *methane* in the domain of Waste Management was created (Table 5).

| **methane (Waste Management)** |
|---|
| A gas whose chemical formula is $CH_4$ produced during the anaerobic decomposition of organic waste in landfills and treatment systems. It is the main component of landfill gas and is generated by methanogenic microorganisms, often alongside carbon dioxide ($CO_2$) as a co-product of microbial processes. Methane emissions contribute significantly to climate change and are more potent than $CO_2$, making their control a key environmental concern. Its production depends on factors such as organic content, water availability, system design, and pH levels. Although methane is a major pollutant, it can also be captured and used as an energy source in waste-to-energy systems. |

Table 5: Definition of *methane* from the point of view of Waste Management

### 5.2 Definition of *methane* in Energy Engineering

In the Energy Engineering corpus, the most frequent contextonyms of *methane* include *gas*, *natural* , *hydrate*, *carbon* and *emission* (Figure 6).

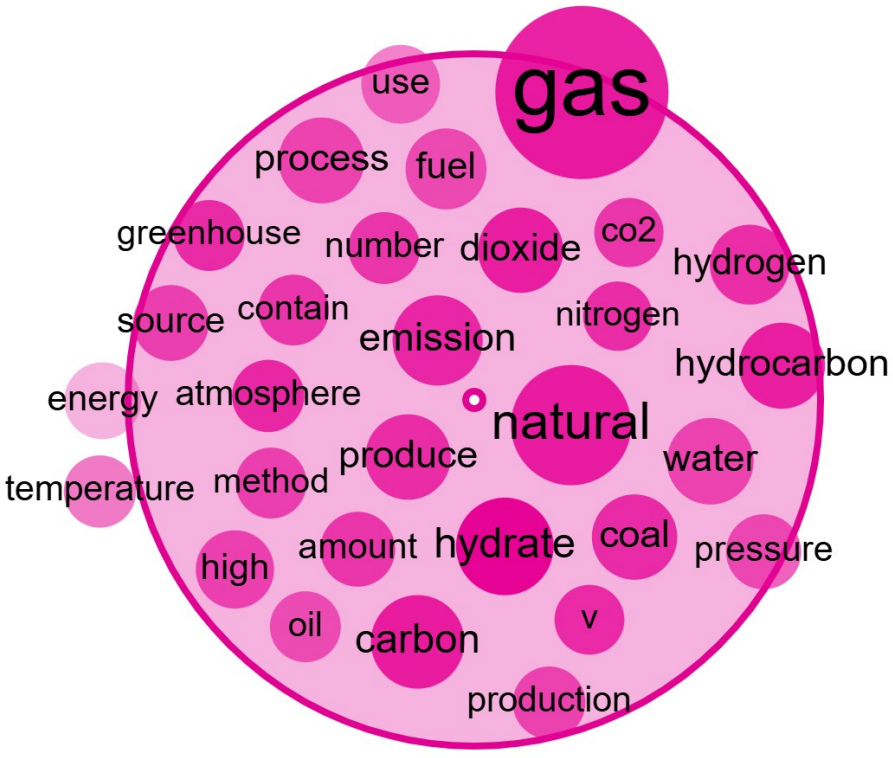


Figure 6: Visualization generated by SkE showing the 30 most frequent contextonyms of *methane* in the Energy Engineering corpus (bubble size indicates frequency)

These contextonyms show that methane is mainly conceptualized as a gas and the main

component of natural gas, widely used as a fuel for energy generation and as a feedstock for hydrogen production. Terms such as *gas*, *natural*, *hydrocarbon*, *fuel*, *energy*, *high*, and *source* emphasize its role as a high-energy compound valued for combustion efficiency. The presence of *hydrate*, *pressure*, and *temperature* highlights its occurrence in methane hydrates, and the physical conditions needed for extraction and storage. Contextonyms like *produce*, *production*, *process*, and *method* reflect the technical focus on how methane is extracted, converted, and used. Meanwhile, terms such as *carbon*, *dioxide*, *$CO_2$*, *emission*, *atmosphere*, and *greenhouse* underscore its climatic impact as a potent greenhouse gas when released into the atmosphere.

Based on the analysis of these contextonyms through the methods explained in the previous sections, a context-sensitive definition of *methane* in the domain of Energy Engineering was created (Table 6).

| **methane (Energy Engineering)** |
| --- |
| A gas whose chemical formula is $CH_4$ that is the main component of natural gas. As a hydrocarbon, it plays a central role as a fuel in energy and hydrogen production and is valued for its high energy content and combustion efficiency. It is obtained from fossil sources such as oil, coal, and methane hydrates (ice-like structures that trap methane in deep ocean sediments and permafrost). Methane is also produced from renewable sources through the anaerobic decomposition of organic waste. Its presence in the atmosphere, whether from controlled use or accidental leakage, contributes significantly to greenhouse gas emissions. For this reason, its containment, control, and optimized production methods are key priorities in minimizing environmental impact while maximizing its role as an energy source. |

Table 6: Definition of *methane* from the point of view of Energy Engineering

## 6. Conclusions

This paper has proposed contextonym analysis as a practical method to support the creation of precise, context-sensitive terminological definitions. Grounded in the FTDA, the method aligns with the view that meaning is shaped by context and that terminologists must identify the most relevant conceptual content depending on contextual constraints. It operationalizes this view by empirically revealing salient features of a term based on its usage in a representative corpus.

Contextonym analysis relies on surface co-occurrence, allowing semantic information to be extracted without predefined relations. It captures a wide range of semantic features that might otherwise go unnoticed. Thanks to the implementation of a custom sketch grammar in SkE, contextonym extraction is now accessible via the WS interface, making the method user-friendly.

A key advantage of contextonym analysis is its interpretability. Each contextonym can be traced back to concordance lines, enabling terminologists to verify results. This contrasts with GenAI tools, whose outputs rely on probabilistic modeling without direct source attribution. Nonetheless, GenAI tools can complement contextonym analysis by quickly suggesting interpretations of contextonyms, which terminologists can then validate through corpus consultation.

This paper has outlined strategies to help terminologists assess which contextonyms are relevant, including filtering techniques, cross-checking with other WS columns, and ranking concordance lines with the GDEX function. It has also emphasized that not all contextonyms are equally useful. Careful selection remains essential.

Ultimately, contextonym analysis empowers terminologists to create context-aware definitions grounded in verifiable data. Combining corpus methods with AI-assisted strategies offers a robust framework for definition writing. This dual approach enhances efficiency and reliability, ensuring that definitions reflect how terms are actually conceptualized in specialized discourse.

## 7. Acknowledgements

This research was funded by Canada's Social Sciences and Humanities Research Council, grant number 430-2023-0248, and Spain's Ministry of Science and Innovation, grant number PID2020-118369GB-I00.

## 8. References

Bowker, L. (2003). Lexical Knowledge Patterns, Semantic Relations, and Language Varieties. *Cataloging & Classification Quarterly*, *37*(1–2), 153–171. https://doi.org/10.1300/J104v37n01_11

Croft, W., & Cruse, A. (2004). *Cognitive Linguistics*. Cambridge University Press.

Drouin, P. (2010). From a Bilingual Transdisciplinary Scientific Lexicon to Bilingual Transdisciplinary Scientific Collocations. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress* (pp. 296–305). Fryske Akademy.

Dubuc, R. (2002). *Manuel pratique de terminologie*. Linguatech.

Evans, V. (2015). A Unified Account of Polysemy Within LCCM Theory. *Lingua*, *157*, 100–123. https://doi.org/10.1016/j.lingua.2014.12.002

Evans, V. (2019). *Cognitive Linguistics: A Complete Guide*. Edinburgh University Press.

Evert, S. (2009). Corpora and Collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook (Volume 2)* (pp. 1212–1248). Mouton de Gruyter.

Fargas, F. X. (2009). *La definició terminològica* (Termcat, Ed.). Eumo.

Freixa, J., & Fernández-Silva, S. (2017). Terminological Variation and the

Unsaturability of Concepts. In P. Drouin, A. Francœur, J. Humbley, & A. Picton (Eds.), *Multiple Perspectives on Terminological Variation* (pp. 155–180). John Benjamins. https://doi.org/10.1075/tlrp.18.07fre

Gadek, G., Betsholtz, J., Pauchet, A., Brunessaux, S., Malandain, N., & Vercouter, L. (2017). Extracting Contextonyms From Twitter for Stance Detection. *ICAART 2017 - Proceedings of the 9th International Conference on Agents and Artificial Intelligence*, *2*, 132–141. https://doi.org/10.5220/0006190901320141

Hanks, P. (2020). How Context Determines Meaning. In G. Corpas Pastor & J.-P. Colson (Eds.), *Computational Phraseology* (pp. 297–310). John Benjamins. https://doi.org/10.1075/ivitra.24.15han

ISO/TC 37/SC 1. (2022). *ISO 704:2022 (Terminology work—Principles and methods)*. ISO.

Jakubíček, M., Měchura, M., Kovář, V., & Rychlý, P. (2018). Practical Post-Editing Lexicography with Lexonomy and Sketch Engine. *XVIII EURALEX International Congress: Lexicography in Global Contexts*. http://euralex2018.cjvt.si/

Ji, H., Ploux, S., & Wehrli, E. (2003). Lexical Knowledge Representation with Contexonyms. *Machine Translation Summit IX*, 194–201.

Kecskes, I. (2023). *The Socio-Cognitive Approach to Communication and Pragmatics*. Springer. https://doi.org/10.1007/978-3-031-30160-5

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten Years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. *Proceedings of the XIII EURALEX International Congress*, 425–432.

Kockaert, H. J., & Steurs, F. (2015). *Handbook of Terminology*. John Benjamins.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(1), 1–31.

León-Araúz, P., & San Martín, A. (2018). The EcoLexicon Semantic Sketch Grammar: From Knowledge Patterns to Word Sketches. In I. Kerneman & S. Krek (Eds.), *LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"* (pp. 94–99). Globalex.

León-Araúz, P., San Martín, A., & Faber, P. (2016). Pattern-Based Word Sketches for the Extraction of Semantic Relations. In P. Drouin, N. Grabar, T. Hamon, K. Kageura, & K. Takeuchi (Eds.), *Proceedings of the 5th International Workshop on Computational Terminology* (pp. 73–82).

Meyer, I. (2001). Extracting Knowledge-Rich Contexts for Terminography. A Conceptual and Methodological Framework. In D. Bourigault, C. Christian, & M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology* (pp. 279–302). John Benjamins.

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In Sojka, Petre & Horák, Aleš (Eds.), *Second Workshop on Recent Advances in Slavonic Natural*

*Language Processing, RASLAN 2008*. Masaryk University.

San Martín, A. (2016). *La representación de la variación contextual mediante definiciones terminológicas flexibles* [PhD Thesis, University of Granada]. https://doi.org/10481/43423

San Martín, A. (2022a). A Flexible Approach to Terminological Definitions: Representing Thematic Variation. *International Journal of Lexicography*, *35*(1), 53–74. https://doi.org/10.1093/ijl/ecab013

San Martín, A. (2022b). Contextual Constraints in Terminological Definitions. *Frontiers in Communication*, *7*. https://doi.org/10.3389/fcomm.2022.885283

San Martín, A. (2024). What Generative Artificial Intelligence Means for Terminological Definitions. In F. Vezzani, G. M. Di Nunzio, B. Sánchez Cárdenas, P. Faber, M. Cabezas García, P. León-Araúz, A. Reimerink, & A. San Martín (Eds.), *3rd International Conference on Multilingual Digital Terminology Today (MDTT 2024)*. CEUR-WS. https://ceur-ws.org/Vol-3703/paper1.pdf

San Martín, A. (2025). Optimizing Contextonymic Analysis for Terminological Definition Writing. *Information*, *16*(4). https://doi.org/10.3390/info16040257

San Martín, A., & Trekker, C. (2021). Adapting Word Sketches for Specialized Knowledge Extraction. In Amalia, Dora, Darnis, Azhari Dasman, Triatna, Amat, & Khairiah, Dewi (Eds.), *14th International Conference of the Asian Association for Lexicography (ASIALEX)* (pp. 64–87). ASIALEX.

San Martín, A., Trekker, C., & Díaz-Bautista, J. C. (2023). Extracting the Agent-Patient Relation from Corpus With Word Sketches. *Proceedings of the 4th Conference on Language, Data and Knowledge*, 666–675. https://aclanthology.org/2023.ldk-1.73.pdf

Seppälä, S. (2015). An Ontological Framework for Modeling the Contents of Definitions. *Terminology*, *21*(1), 23–50. https://doi.org/10.1075/term.21.1.02sep

Șerban, O., Pauchet, A., Rogozan, A., & Pécuchet, J.-P. (2012). Semantic Propagation on Contextonyms Using SentiWordNet. *WACAI 2012, Workshop Affect, Compagnon Artificiel, Interaction*, 86–94.

Suonuuti, H. (1997). *Guide to Terminology*. Tekniikan Sanastokeskus ry. http://www.nordterm.net/wiki/en/index.php/Nordterm_8

Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Sociocognitive Approach*. John Benjamins.

Vézina, R., Darras, X., Bédard, J., & Lapointe-Giguère, M. (2009). *La rédaction de définitions terminologiques*. Office québécois de la langue française.