Extracting the Agent-Patient Relation From Corpus With Word Sketches

Antonio San Martín and **Catherine Trekker**
University of Quebec in Trois-Rivières
Trois-Rivières, Canada
{sanmarti,trekkers}@uqtr.ca

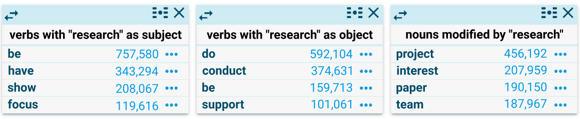
Juan Carlos Díaz-Bautista
Autonomous Mexico State University
Toluca, Mexico
jdiazb002@alumno.uaemex.mx

Abstract

Word sketches are a powerful function of Sketch Engine that automatically summarizes the most common usage patterns of a search word in a corpus. While they have proven to be a valuable tool for collocational analysis in both general and specialized language, their potential for the extraction of terminological knowledge is yet to be fully realized. To address this, we introduce a novel semantic sketch grammar designed to extract the agent-patient relation, an important yet understudied relation. This paper presents the various stages of developing the rules that compose this sketch grammar as well as the evaluation of their precision. The errors identified during the evaluation process are also analyzed to guide future improvements. The sketch grammar is available online so that any user can apply it to their own corpora in Sketch Engine.

1 Introduction

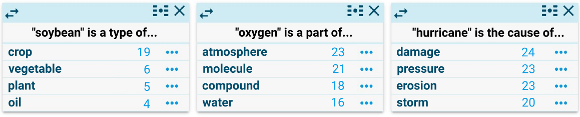
Word sketches (WSs) are a powerful function of corpus analysis tool Sketch Engine (https://www.sketchengine.eu/) (Kilgarriff et al., 2014) that automatically summarizes the most common usage patterns of a search word in a corpus. A WS is composed of columns listing the words that are related (most often syntactically) to the search word in the corpus. This includes, for instance, the verbs having the search word as subject or object, or the words modified by the search word (Figure 1). WSs have proven valuable for collocational analysis in both general and specialized language, as they enable the easy identification of a word's combinatorial behavior.



verbs with "research" as subject		verbs with "research" as object		nouns modified by "research"	
be	757,580	do	592,104	project	456,192
have	343,294	conduct	374,631	interest	207,959
show	208,067	be	159,713	paper	190,150
focus	119,616	support	101,061	team	187,967

Figure 1: Three WS columns of the search word *research* in the enTenTen21 corpus

However, the default WS in Sketch Engine is not adapted to the extraction of terminological knowledge. For this reason, the EcoLexicon Semantic Sketch Grammar (ESSG) (León-Araúz et al., 2016; León-Araúz and San Martín, 2018; San Martín et al., 2022) expanded WS functionality to enable the identification of some of the most common relations used in Terminology and Ontology Engineering with new WS columns (generic-specific, part-whole, cause, function, and location) in English and French (Figure 2).



"soybean" is a type of...		"oxygen" is a part of...		"hurricane" is the cause of...	
crop	19	atmosphere	23	damage	24
vegetable	6	molecule	21	pressure	23
plant	5	compound	18	erosion	23
oil	4	water	16	storm	20

Figure 2: Semantic WS columns generated with the ESSG in the EcoLexicon English Corpus (León-Araúz and San Martín, 2018)

This paper presents the first version of a novel semantic sketch grammar designed to extract the agent-patient relation in the form of WSs. An example of this relation is the one between *mechanic* and *tire* in "...the mechanic inflated the tires...", "...mechanics mount tires..." and "...the tires were balanced by a mechanic...". In all three examples, *mechanic* is the agent of the action that affects *tire*, which is the patient (*mechanic* affects *tire*)[1].

The agent-patient is a valuable relation for the extraction and representation of terminological knowledge because the organization of specialized domains is shaped by the interaction between different agents and patients (Faber, 2015). Despite its importance, it is an understudied relation, and terminologists and ontologists currently lack a straightforward way of extracting it from corpora. Our proposal seeks to bridge this gap by providing

[1] Inspired by the "affects" relation in EcoLexicon (León-Araúz and Faber, 2010), a terminological knowledge base on the environment, we will use the verb *affect* to represent the agent-patient relation in a triplet.

a solution for extracting this semantic relation in the form of WSs. By facilitating the analysis of the interplay of agents and patients within specialized domains, this tool can contribute to both practical terminological and ontological work and academic research.

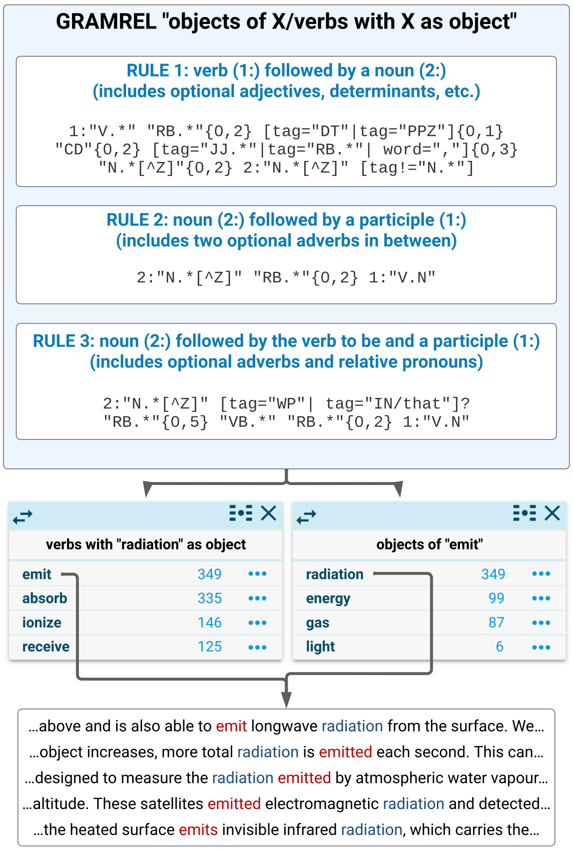
The remaining sections of this paper are structured as follows. Section 2 describes the process of WS generation. In Section 3, we present our definition of agent, patient, and the agent-patient relation. Section 4 introduces the methods and materials employed in developing the new agent-patient sketch grammar. Sections 5 and 6 outline the two main development phases. The evaluation results are discussed in Section 7. Finally, Section 8 gives the conclusions derived from this research and outlines future work.

2 Word Sketch Generation

WS generation in Sketch Engine is based on the matching of patterns encoded as rules expressed in CQL language (Jakubíček et al., 2010). A CQL rule is composed of tokens in the form of attributes (part-of-speech tag, lemma, word form, etc.) and values combined with regular expressions. For example, the rule `[tag="J.*"] [tag="N.*"] [lemma="management"]` matches concordances containing the lemma *management* preceded by a noun and an adjective (e.g., “natural resource management”, “effective risk management”, and “cold chain management”).

Within a CQL rule intended for WSs, the position of the words to be extracted as the WS results are identified. For instance, the rule `1:[tag="J.*"] [tag="J.*"]? 2:[tag="N.*"]` enables the extraction of an adjective (1:) that is followed by another optional adjective and a noun (2:). It also allows the inverse: the extraction of a noun (2:) preceded by an optional adjective, which itself is preceded by another adjective (1:). In this case, Sketch Engine identifies matches of the rule (a noun preceded by one or two adjectives) in the corpus, and subsequently extracts the left-most adjective and the noun from each matched concordance. However, a significant limitation of WSs is that results are restricted to single words.

For WS generation, the CQL rules designed to identify the same relation are grouped into a gramrel (for “grammatical relation”). Each gramrel can produce one or more WS columns (normally one relation and its reverse). The collection of gramrels that generate a WS is referred to as a sketch grammar. For instance, the gramrel included in Sketch Engine’s default sketch grammar that identifies the relation between the object of a sentence and its verb generates two WS columns ("objects of “X”" and its reverse "verbs with “X” as object") by means of three rules (Figure 3). The first rule identifies the object-verb relation in the active voice and the other two in the passive voice (one without the verb *to be* and the other with it).



GRAMREL "objects of X/verbs with X as object"

RULE 1: verb (1:) followed by a noun (2:)
(includes optional adjectives, determinants, etc.)

```
1:"V.*" "RB.*"{0,2} [tag="DT"|tag="PPZ"]{0,1}
"CD"{0,2} [tag="JJ.*"|tag="RB.*"| word=","]{0,3}
"N.*[^Z]"{0,2} 2:"N.*[^Z]" [tag!="N.*"]
```

RULE 2: noun (2:) followed by a participle (1:)
(includes two optional adverbs in between)

```
2:"N.*[^Z]" "RB.*"{0,2} 1:"V.N"
```

RULE 3: noun (2:) followed by the verb to be and a participle (1:)
(includes optional adverbs and relative pronouns)

```
2:"N.*[^Z]" [tag="WP"| tag="IN/that"]?
"RB.*"{0,5} "VB.*" "RB.*"{0,2} 1:"V.N"
```

verbs with "radiation" as object		
emit	349	•••
absorb	335	•••
ionize	146	•••
receive	125	•••

objects of "emit"		
radiation	349	•••
energy	99	•••
gas	87	•••
light	6	•••

…above and is also able to emit longwave radiation from the surface. We…
…object increases, more total radiation is emitted each second. This can…
…designed to measure the radiation emitted by atmospheric water vapour…
…altitude. These satellites emitted electromagnetic radiation and detected…
…the heated surface emits invisible infrared radiation, which carries the…

Figure 3: The "objects of “X”/verbs with “X” as object" gramrel in the default English sketch grammar with an example from the EcoLexicon English Corpus

While the default sketch grammar is mainly based on syntactic relations, the ESSG extracts semantic relations by means of knowledge patterns, i.e., lexico-syntactic patterns that match contexts where a specific semantic relation is conveyed (Meyer, 2001). For instance, the knowledge pattern “X and other Y” (e.g., “…theophylline and other bronchodilators…”) conveys a generic-specific relation (*theophylline* is-a *bronchodilator*).

While our new agent-patient sketch grammar extracts a semantic relation, our methodology does not rely on knowledge patterns[2]. Instead, our starting point is the syntactic relation between the nouns functioning as subject and object in the same sentence. This is based on the premise that the subject typically functions as the agent and the object as a patient. Even though the subject-object relation does not always correspond to an agent-patient semantic relation (and vice versa), the results of a pilot study confirmed the feasibility of this approach (San Martín and Trekker, 2021).

3 Defining the Agent-Patient Relation

We define the agent-patient relation as one in which one participant in the action (the agent) affects another participant (the patient) in some way. In this sense, we adopt the notions of agent and patient in a broad sense, aligning with Dowty's (1991) macroroles of proto-agent and proto-patient, or Van Valin's (2004) actor and undergoer. This implies that our definition of agent also encompasses other semantic roles that affect another participant in the action such as effector, actor, instrument, and others. Similarly, our interpretation of patient is inclusive of roles that other authors might label not only as patient but also as theme, referent, goal, beneficiary, result, etc. As a result, according to our definition, agents and patients can be nouns that refer to any type of concept including concrete and abstract entities, processes, states, and attributes.

The extent to which an agent's action must impact a patient in order to establish the existence of an agent-patient relation is not clear-cut. Whereas "...the researcher vaccinated the rats..." is indisputably agentive and "...the researcher imagined colorful rats...", non-agentive, there are many borderline cases, such as "...the researcher possesses rats..." or "...the researcher exhibits the rats...".

To better delimit the agent-patient relation for the creation and subsequent evaluation of CQL rules, we used a pre-existing list of verb senses to determine which ones are to be considered agentive and which are not. We chose that of Faber and Mairal Usón (1999), which classifies the English verb lexicon into 13 verb sense categories (such as existence, movement, and position), which are further subdivided into 389 subcategories.

We labeled each verb sense in the list as agentive, non-agentive, or intransitive, based on their nature. Given the fuzziness of the agent-patient relation, there were unavoidably subjective choices. Most verb senses were deemed either agentive or intransitive. Agentive subcategories include, among others, all causative senses, which means that our definition of the agent-patient relation subsumes the causal relation. Intransitive subcategories are those involving a single argument.

The non-agentive subcategories included those verb senses overlapping with the part-whole and location relations. Additionally, other subcategories that were considered non-agentive include, among others, those expressing perception, cognition, feeling, and speech. Some possession verb senses were also considered non-agentive, such as those expressing basic possession (*have*, *possess*, *own*). However, when the agent carries out an action to possess something (*take*, *get*, *obtain*) or there is a transfer of possession (*give*, *provide*, *exchange*), the verb senses are considered agentive. The final classification of verb senses is available at http://doi.org/10.5281/zenodo.8121939[3].

As will be seen later, verbs that most frequently activate intransitive or non-agentive senses were filtered out in the CQL rules.

4 Materials and Methods

The development of a new sketch grammar is based on the encoding of CQL rules and their subsequent enhancement based on the evaluation of the matching concordances in a given corpus (León-Araúz et al., 2016). For this agent-patient sketch grammar (consisting of a single gramrel)[4], we used the Elsevier OA CC-BY Corpus (Kershaw and Koeling, 2020), which is composed of 40,000 open-access articles in English published between 2014 and 2020 in Elsevier journals. The corpus in its version available in Sketch Engine contains 187,615,459 words and 232,511,611 tokens. It covers a wide variety of domains (e.g., Medicine, Computer Science, Social Sciences, Economics, Arts, etc.). This ensures that the sketch grammar is domain-independent.

[2] However, some of the CQL rules, as will be seen below, could be considered knowledge patterns.

[3] In this URL, the final sketch grammar can also be found, as well as all the lists of verbs and phrases used to build the CQL rules that are mentioned later in the paper.

[4] In San Martín and Trekker (2021), we created a preliminary version of this gramrel. The one presented in this study partly follows the same methodology, but with numerous improvements and modifications. These differences cannot be discussed because of space restrictions.

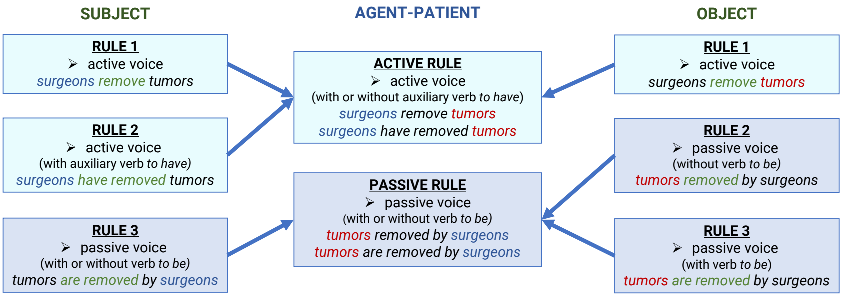

SUBJECT
RULE 1
➢ active voice
surgeons remove tumors
RULE 2
➢ active voice
(with auxiliary verb to have)
surgeons have removed tumors
RULE 3
➢ passive voice
(with or without verb to be)
tumors are removed by surgeons
AGENT-PATIENT
ACTIVE RULE
➢ active voice
(with or without auxiliary verb to have)
surgeons remove tumors
surgeons have removed tumors
PASSIVE RULE
➢ passive voice
(with or without verb to be)
tumors removed by surgeons
tumors are removed by surgeons
OBJECT
RULE 1
➢ active voice
surgeons remove tumors
RULE 2
➢ passive voice
(without verb to be)
tumors removed by surgeons
RULE 3
➢ passive voice
(with verb to be)
tumors are removed by surgeons

Figure 4: Generation of the simple version of the agent-patient rules

Our initial step was to generate a simple version of the gramrel by integrating the two default gramrels "objects of “X”/verbs with “X” as object" (object gramrel) and "subjects of “X”/verbs with “X” as subject" (subject gramrel) (Figure 4). The active-voice rules were combined into a new rule (‘active-simple’), while the passive ones were also consolidated into another one (‘passive-simple’).

We then proceeded to the subject-object enhancement, which consisted of enriching and refining the simple version to improve its precision and recall with respect to the extraction of the subject-object relation. This was followed by the agent-patient enhancement, aimed at improving its capacity to extract the agent-patient relation.

Throughout both enhancement phases, minor and major evaluations were carried out, with the authors of the paper acting as evaluators. All evaluations were collaboratively reviewed and agreed upon, aimed at iteratively refining the rules, determining whether 20 random concordances extracted with the evaluated rule conveyed the subject-object relation or the agent-patient relation (depending on the enhancement phase). For a concordance to be considered valid, the rule also had to correctly identify the nouns functioning as subject and object (or agent and patient) within the concordance.

The count of valid concordances was used to estimate precision and determine whether the evaluated modifications should be retained. When the results were inconclusive, additional sets of 20 concordances were evaluated. Recall was prioritized over precision since users ultimately access the results of the gramrel through WSs, where the potentially most relevant results (with higher frequency) are at the top of the WS column.

In this paper, we only present the results of the major evaluations which involved the assessment of 250 random concordances and were reserved for definitive versions of the rules.

5 Subject-Object Enhancement

For the subject-object enhancement phase, the rules resulting from combining the subject and object gramrels (‘active-simple’ and ‘passive-simple’) were enriched and refined to increase recall without compromising precision. Each enrichment was subject to a minor evaluation. These enhancements included, among others, the addition of optional modal and auxiliary verbs, the possibility of more than one main verb, optional gerunds and participles where adjectives were already possible, an optional comma before the optional relative pronoun as well as some minor adjustments to avoid noise (for instance, excluding the presence of *so* before the optional relative pronoun to avoid noise created by the occurrence of *so that*).

Both versions of the rules were subject to a major evaluation. For a concordance to be considered valid, there needs to be a subject-object relation between the identified nouns, and both of them need to be the head of their noun phrase.

The evaluation results (Figure 5) indicate that the simple and enhanced versions yield comparable subject-object precision. However, the enhanced active rule extracts 53.74% more concordances, and the enhanced passive rule extracts 31.86% more concordances than their simple counterparts.

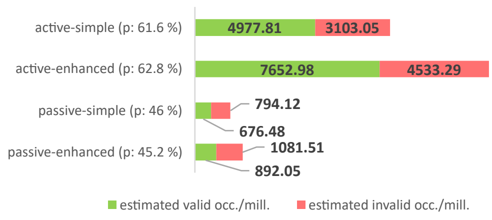


Figure 5: Precision and occurrences per million tokens of the simple and enhanced rules

6 Agent-Patient Enhancement

Since the two enhanced rules provided a precision comparable to the simple ones but with higher recall, the agent-patient enhancement was performed on these two rules. However, before proceeding, an evaluation of the agent-patient precision of the same concordances was performed to establish a reliable baseline.

Evaluators answered the following question for each concordance: "Does the identified agent have an effect on the identified patient?". When the concordance was not considered valid, the error or errors at cause were noted. Although an agent-patient relationship is established in the concordance, if the correct agent and patient are not identified, the concordance is considered invalid. The list of errors and their distribution in this evaluation and the subsequent ones are reproduced and explained in section 7.2.

According to the results of the evaluation (Figure 6), 'active-enhanced' has an agent-patient precision of 31.2% and 'passive-enhanced', 38.4%. Both values are significantly lower than their subject-object precision. This indicates that solely focusing on improving subject-object precision is insufficient for effectively capturing the agent-patient relation. Consequently, we proceeded to the agent-patient enhancement, which was divided into three stages described in the remainder of this section.

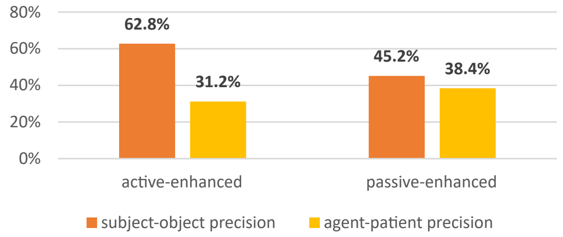


Figure 6: Evaluation results of 'active-enhanced' and 'passive-enhanced'

6.1 First Stage

This first stage, aimed at improving precision[5], consisted of creating a version of the rules where verbs that do not convey the agent-patient relation are excluded. To compile a list of non-agentive verbs, we first extracted the 1000 most frequent verbs in the Elsevier corpus as well as the 1000 most frequent verbs in the same corpus occurring within our active and passive enhanced rules. The elimination of duplicates produced a list of 1083 verbs, which was reduced to 1054 verbs after the consolidation of spelling variants and lemmatization errors.

Each verb was subjected to a minor evaluation in which its presence was forced in the active and passive rules. The purpose of the evaluation was to determine whether the verb more frequently activates agentive or non-agentive verb senses, based on our classification of verb senses.

Since verbs can have both agentive and non-agentive senses because of polysemy, verbs with non-agentive senses in 75% or more of the concordances were classified as non-agentive. As a result, a total of 275 non-agentive verbs (e.g., *say*, *define*, *display*...) were identified, as well as 693 agentive verbs (e.g., *convert*, *target*, *structure*...).

We also identified intransitive and inverting verbs. Intransitive verbs produce noise because they cannot instantiate an agent-patient relation. An intransitive verb is one that in 75% or more of the concordances was found to be intransitive. A total of 76 intransitive verbs were thus identified (e.g., *exist*, *go*, *live*...).

As for inverting verbs, they are verbs in which the subject functions as the patient and the object as the agent. For instance, *undergo* in "...women undergo an outpatient hysteroscopy..." (hysteroscopy *affects* woman). We identified 10 inverting verbs (e.g., *experience*, *resist*, *tolerate*...).

With the final list of verbs, we created four variants of the rules. The first two rules ('active-exc' and 'passive-exc') exclude non-agentive, intransitive, and inverting verbs[6]. Conversely, the other two rules ('active-inv' and 'passive-inv') only permit inverting verbs and reverse the order in which the agent and the patient appear.

6.2 Second Stage

The second stage, aimed at improving recall, consisted of the creation of a version of the active rule that allows certain prepositional verbs[7] that convey an agent-patient relation (e.g., *lead to*, *contribute to*, *aim at*, *help in*). A version of the passive rule that permits certain verbs followed by prepositions other than *by* was also created (e.g., *attribute to*, *expose to*, *filter through*).

[5] Henceforth, precision is understood specifically as agent-patient precision.

[6] The gerund verb forms *using* in 'active-exc' and *facing* in 'active-inv' were excluded too because they generated excessive noise.

[7] By prepositional verbs, we also mean particle verbs.

For the active rule ('active-prep'), we initially allowed the optional presence of a preposition or a particle after the main verb. However, the evaluation of the concordances of 26 prepositions and particles in that position showed that this approach created a significant amount of noise. Nonetheless, this evaluation allowed us to identify 148 prepositional verbs that could potentially be agentive.

After an individual evaluation of each one, the list was reduced to 107 agentive prepositional verbs (e.g., *act on*, *contribute to* or *deal with*[8]). This permitted the creation of the rule 'active-prep'. Also identified were 16 inverting prepositional verbs (e.g., *suffer from*, *depend on* or *result from*), resulting in the rule 'active-prep-inv'.

Some examples of valid concordances from these two rules include "...Government can contribute to realising a circular economy..." (*government* affects *economy*) and "...mice reacted to fear conditioning stimuli..." (*stimulus* affects *mouse*).

Using this method and by means of minor iterative evaluations, we identified three verbs that can appear in passive voice without a by-phrase but which are followed by a prepositional phrase with agentive meaning: *attributed to*, *exposed to* and *filtered through*. The rule 'passive-prep' forces their presence.

Some examples of valid concordances retrieved with this rule include "...Supernatants were filtered through a 0.45 μm membrane..." (*membrane* affects *supernatant*) and "...sorption could therefore be attributed to the sludge..." (*sludge* affects *sorption*).

6.3 Third Stage

Finally, the third stage, also aimed at improving recall, consisted of developing a version of the active rule that allows verb phrases expressing an agent-patient relation (e.g., *to have impact/effect/influence on*, *to play a role in*, *to make a contribution to*...). Additionally, we created a version of the passive rule where *by* is replaced by expressions such as *using*, *by means of*, *with the help of*, etc. (e.g., "...rules are instituted with the help of a dietician...").

In the case of verb phrases, the patient is not the object of the sentence but rather the head of the prepositional phrase that follows. For instance, in "competition has a sizeable negative impact on pupil wellbeing", *wellbeing* serves as the patient despite not being the object. Considering this, we developed a version of the active rule ('active-phrases') that forces the presence of agentive verb phrases such as *play a role in*, *have effect on* or *make use of* and retrieves as patient the head of the prepositional phrase that follows.

Each verb phrase was individually evaluated to ensure a minimum precision level of 50%. An example of valid concordances extracted with this rule are "...Mitochondria play key roles in mammalian apoptosis..." (*mitochondrion* affects *apoptosis*) and "...Imports have large positive effects on firm productivity..." (*import* affects *productivity*).

Additionally, we created a passive rule ('passive-not-by') where the by-phrase is replaced by expressions referring to an instrument or a means such as *using*, *by means of*, and other variants. Each of the expressions in the rule was evaluated to determine whether they provided at least 50% precision. An example of some valid concordances extracted with this rule are "...The pycnometer was calibrated using a standard calibration ball..." (*ball* affects *pycnometer*) and "...sequences can be folded by addition of metal ions..." (*ion* affects *sequence*).

7 Evaluation Results

7.1 Precision

Figure 7 presents the results of the evaluation of each of the rules that make up the new agent-patient gramrel. The figure also includes the number of valid matches that each rule is estimated to retrieve from the Elsevier corpus (expressed as occurrences per million tokens). This estimate was calculated by applying the precision percentage to the total number of matches retrieved by each rule.

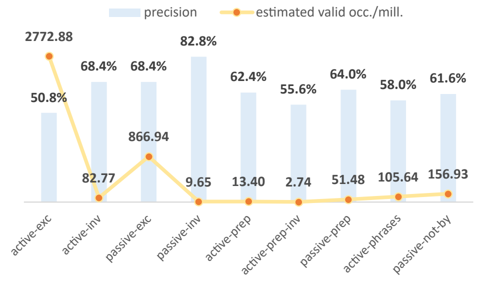


Figure 7: Evaluation results per rule

[8] The gerund of *deal* (i.e., *dealing*) was excluded from the rule because, unlike other tenses, it mostly had a non-agentive sense.

With an overall precision of 54.9%, the new gramrel significantly outperforms the baseline (32.2%) (Figure 8). Each individual rule also surpasses the baseline in precision. However, the total count of valid occurrences per million tokens retrieved by the gramrel is slightly lower than the baseline, although the number of invalid matches (i.e., noise) is nearly three times lower.

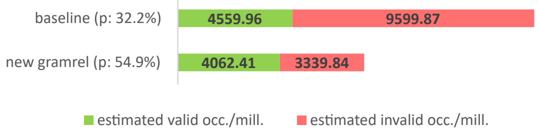


Figure 8: Precision and occurrences per million tokens of the baseline and the new gramrel

Nearly 90% of the valid occurrences recovered by the new gramrel are attributed to two rules: ‘active-exc’ and ‘passive-exc’, which capture the subject-object relation but block selected verbs. Passive rules also exhibit more precision than active rules because of their inherent restrictiveness. Unlike the flexibility in verb tense allowed by active rules, passive rules need the presence of a past participle, which mitigates potential noise.

It is worth noting that whereas assessing rule precision through random concordances is useful during the development process, only the analysis of the resulting WS can validate the usefulness of the sketch grammar. Terms unlikely to be queried by a user through the WS function (due to their irrelevance in terminological analysis or because they do not engage in agent-patient relations) are identified as potential agents or patients in these random concordances. Consequently, random concordances tend to be noisier than those associated with genuine WS queries made by terminologists or ontologists. Moreover, WSs show the most frequent results at the top, which tend to be linked to a higher number of valid concordances.

Since this agent-patient sketch grammar is still in development and WS evaluation is a labor-intensive task, the resulting WSs will only be evaluated when the final version is completed.

7.2 Types of Errors

The following six types of errors were identified during the evaluation:

1. *Non-agentive*: The relation between the two nouns is not agent-patient because the verb sense is non-agentive (e.g., “...results indicate a temperature increase...”). Evaluators referred to the verb sense classification to determine the agentivity of the verb sense within each concordance. The *non-agentive* error also includes the cases in which the agent was erroneously retrieved as a patient and vice versa. For example, in “...drivers experiencing more fatigue...”, the correct relation is “*fatigue* affects *driver*” and the inverse would be considered an error under this category.

2. *Not head*: The retrieved noun is not the head of the grammatical subject or object. This can be caused by multiword terms, prepositional phrases, relative clauses, etc. For instance, in “...The discharge of untreated or partially treated domestic wastewater to the aquatic environment severely threatens public health...”, *environment* was mistakenly detected as the agent instead of *discharge*.

   When the agent or patient is a noun phrase, it may be unclear which is the most semantically significant noun. To ensure objectivity, we followed a strict syntactic criterion with a short list of exceptions such as *group of*, *part of*, etc., where it was determined that the correct noun is not the head. For instance, in “...A number of researchers have used salt...”, although *researchers* is not the head, it was considered a valid concordance.

3. *Not noun*: A noun that is not the subject or object is retrieved because the subject or object is not a noun phrase, but rather a clause or a pronoun (e.g., “...Understanding how meteorology impacts the seasonality of Lyme disease case occurrence can aid in targeting limited prevention resources...”). This type of error also includes cases where an incorrect noun is retrieved as agent because the subject is not explicit in the sentence (e.g., “...Accelerometers are glued to the surface of the plate using hot glue...”).

4. *POS tagging*: Due to a POS tagging error, an incorrect agent-patient relation is retrieved. For instance, the concordance “...the total number generated matches the distribution of the dwelling stock...” was incorrectly retrieved because *matches* was tagged as a noun instead of a verb.

5. *Not by-phrase*: For passive rules, the noun that follows the preposition *by* is not the logical subject. For instance, in "...This enables dry commodities to be marketed by weight...", weight is not the passive logical subject, but the head of an adverbial. Nonetheless, in those cases in which the adverbial headed by *by* introduces an instrument or a means, they were considered valid. For instance, in "...the tissue had already been stabilised by fixation...", although *fixation* is not the logical subject, the concordance was considered valid (*fixation* affects *tissue*).

6. *Segmentation*: An invalid agent-patient relation is retrieved due to a segmentation error (e.g., "...and to extract B. Exponentially growing cells were...").

Figure 9 illustrates the distribution of error types per rule. Since a single concordance can contain more than one type of error, the count of errors may not match the number of invalid concordances (out of 250 evaluated concordances per rule).

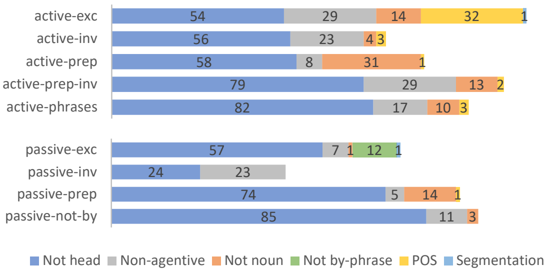


Figure 9: Distribution of error types per rule

The *not head* error accounts for over half of the errors in all rules. This error is a byproduct of the fact that WSs can only extract one-word results.

The way our rules select which noun to identify as agent or patient is inherited from how it is done in Sketch Engine's default sketch grammar. Before the verb, the rules capture the rightmost noun and, after the verb, the rightmost noun before any non-noun token. This approach yields precise results in the absence of prepositional phrases (e.g., "...energy suppliers use wastewater heat to produce...").

However, the presence of prepositional phrases before the verb is the cause of a considerable amount of noise (e.g., "...Hydrodynamics in bubble columns strongly influence mass transfer..."). In fact, the difference in the number of *not head* errors between rules can be primarily attributed to the varying frequency of prepositional phrases occurring before the verb in each rule.

As for the *POS tagging* error, it is significantly more prevalent in the 'active-exc' rule because of the POS tagger's difficulty in distinguishing between past tense verbs and past participles (e.g., "...there is growing evidence that increased production and productivity can lead...") as well as present participles and nouns (e.g., "...solar absorption cooling system...").

In 'active-prep', we found more *not noun* errors than in other rules because some of the prepositional verbs included in the rule have a greater tendency to have a clause as subject, notably *lead to* and *contribute to* (e.g., "...Increasing the amount of rutile phase compared to that of the anatase phase led to decrease the photodegradation...").

Finally, the *not by-phrase* error is exclusive to 'passive-exc' and 'passive-inv' because the other passive rules do not match concordances with by-phrases. However, in 'passive-inv', we did not find this error because the inverting verbs allowed by this rule do not normally induce this error.

7.3 Avenues of Improvement

The evaluation of the rules has underscored the priorities to be addressed for the development of the final version of the sketch grammar.

The fact that most concordances retrieved by the gramrel are extracted by the 'active-exc' and 'passive-exc' rules suggests that future improvement efforts should focus on increasing the precision of these two rules. One way to accomplish this would be to limit the retrieval as a patient of the object of common verb phrases. For instance, the rule 'active-exc' currently retrieves non-agentive concordances such as "...30% of cycling takes place in roads..." or "...data may shed light on HBP dysfunction...". These noisy concordances could be excluded by not allowing *place* and *light* as patient when their respective verbs are *take* and *shed*.

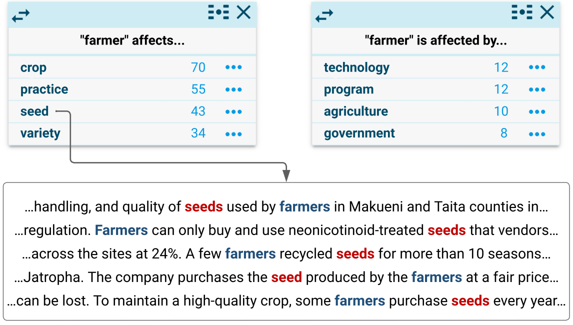
Still another possibility is the expansion of our list of non-agentive, intransitive, and inverting verbs, which are specifically excluded in 'active-exc' and 'passive-exc'.

Finally, considering that the *not head* error accounts for over half of all errors across all rules, it could be productive to examine how different types of multiword terms in the agent or patient position, as well as the presence of prepositional phrases, can be accounted for in the rules.

8 Conclusions and Future Work

In this paper, we have presented the development of an innovative sketch grammar that enables users to extract the agent-patient relation from any English user-owned corpus in Sketch Engine. The current version of the agent-patient sketch grammar can be downloaded at http://doi.org/10.5281/zenodo.8121939, where instructions on how to use it with their own corpora in Sketch Engine are also found.

Figure 10 shows a sample of the resulting agent-patient WS columns for the term *farmer* when the sketch grammar is applied to an 8-million-word specialized corpus on agriculture. Some of the concordances that are accessible via the WS are also reproduced.



"farmer" affects...		
crop	70	•••
practice	55	•••
seed	43	•••
variety	34	•••

"farmer" is affected by...		
technology	12	•••
program	12	•••
agriculture	10	•••
government	8	•••

...handling, and quality of **seeds** used by **farmers** in Makueni and Taita counties in...
...regulation. **Farmers** can only buy and use neonicotinoid-treated **seeds** that vendors...
...across the sites at 24%. A few **farmers** recycled **seeds** for more than 10 seasons...
...Jatropha. The company purchases the **seed** produced by the **farmers** at a fair price...
...can be lost. To maintain a high-quality crop, some **farmers** purchase **seeds** every year...

Figure 10: Agent-patient WS columns of *farmer* in an agricultural corpus

The current agent-patient sketch grammar, though currently functional, is still under development and will undergo future enhancements to increase both precision and recall, including those previously mentioned in this paper. As with the current version, subsequent iterations will be made freely accessible online.

The agent-patient sketch grammar can greatly benefit terminologists and ontologists since it facilitates access to one aspect that reflects how specialized domains are structured that was previously very time-consuming to extract. Beyond its practical applications, this sketch grammar is a valuable research tool. We plan to use it in future studies to further explore the agent-patient relation in specialized domains.

Acknowledgements

This research was carried out as part of projects 2020-NP-267503 funded by Quebec's Society and Culture Research Fund, 430-2023-0248 funded by the Social Sciences and Humanities Research Council of Canada, PID2020-118369GBI00 funded by the Spanish Ministry of Science and Innovation, and A-HUM-600-UGR20 funded by the Regional Government of Andalusia.

References

David Dowty. 1991. Thematic Proto-Roles and Argument Selection. *Language*, 67(3):547–619.

Pamela Faber. 2015. Frames as a Framework for Terminology. In H. J. Kockaert and F. Steurs, editors, *Handbook of Terminology, volume 1*, pages 13–33. John Benjamins, Amsterdam.

Pamela Faber and Ricardo Mairal Usón. 1999. *Constructing a Lexicon of English Verbs*. Mouton de Gruyter, Berlin.

Miloš Jakubíček, Adam Kilgarriff, Diana McCarthy, and Pavel Rychlý. 2010. Fast syntactic searching in very large corpora for many languages. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 741–747, Tohoku University, Sendai, Japan. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Daniel Kershaw and Rob Koeling. 2020. Elsevier OA CC-BY Corpus.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Pilar León-Araúz and Pamela Faber. 2010. Natural and contextual constraints for domain-specific relations. In *The Workshop Semantic Relations, Theory and Applications*, pages 12–17, Valletta.

Pilar León-Araúz and Antonio San Martín. 2018. The EcoLexicon Semantic Sketch Grammar: from Knowledge Patterns to Word Sketches. In *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"*, pages 94–99, Miyazaki. Globalex.

Pilar León-Araúz, Antonio San Martín, and Pamela Faber. 2016. Pattern-based word sketches for the extraction of semantic relations. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, pages 73–82, Osaka, Japan. The COLING 2016 Organizing Committee.

Ingrid Meyer. 2001. Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In *Recent Advances in Computational Terminology*, pages 279–302. John Benjamins, Amsterdam.

Antonio San Martín and Catherine Trekker. 2021. Adapting word sketches for specialized knowledge extraction. In *Proceedings of the 14th International Conference of the Asian Association for Lexicography (ASIALEX)*, pages 64–87, Jakarta. ASIALEX.

Antonio San Martín, Catherine Trekker, and Pilar León-Araúz. 2022. Repérage automatisé de l'hyponymie dans des corpus spécialisés en français à l'aide de Sketch Engine. *Terminology*, 28(2):264–298.

Robert D. Van Valin. 2004. Semantic macroroles in Role and Reference Grammar. In R. Kailuweit and M. Hummel, editors, *Semantische Rollen*, pages 62–82. Narr, Tübingen.