Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5953–5961
Marseille, 11–16 May 2020

Extraction of Hyponymic Relations in French with Knowledge-Pattern-Based Word Sketches

Antonio San Martín, Catherine Trekker, Pilar León-Araúz
University of Quebec in Trois-Rivières, University of Granada
3351, boul. des Forges, Trois-Rivières (Quebec, Canada), C/ Buensuceso, 11, Granada (Spain)
antonio.san.martin.pizarro@uqtr.ca, catherine.trekker-seguin@uqtr.ca, pleon@ugr.es

Abstract

Hyponymy is the cornerstone of taxonomies and concept hierarchies. However, the extraction of hypernym-hyponym pairs from a corpus can be time-consuming, and reconstructing the hierarchical network of a domain is often an extremely complex process. This paper presents the development and evaluation of the French EcoLexicon Semantic Sketch Grammar (ESSG-fr), a French hyponymic sketch grammar for Sketch Engine based on knowledge patterns. It offers a user-friendly way of extracting hyponymic pairs in the form of word sketches in any user-owned corpus. The ESSG-fr contains three times more hyponymic patterns than its English counterpart and has been tested in a multidisciplinary corpus. It is thus expected to be domain-independent. Moreover, the following methodological innovations have been included in its development: (1) use of English hyponymic patterns in a parallel corpus to find new French patterns; (2) automatic inclusion of the results of the Sketch Engine thesaurus to find new variants of the patterns. As for its evaluation, the ESSG-fr returns 70% valid hyperonyms and hyponyms, measured on 180 extracted pairs of terms in three different domains.

Keywords: Corpus (Creation, Annotation, etc.), Information Extraction, Information Retrieval, Knowledge Discovery/Representation, Semantics, Tools, Systems, Applications, Validation of LRs

1. Introduction

Semantic relations are one of the main data categories included in terminological knowledge bases (TKBs) since they structure specialized knowledge and are useful both for final users and terminographers. It is widely known that users acquire knowledge through the visualization of concept systems and are thus able to expand previously stored knowledge. In contrast, terminographers benefit from the study of semantic relations while compiling other related data, such as concept systems, category templates, definitions, and contexts. Accordingly, hyponymy is a crucial relation in Terminology since it is the cornerstone of taxonomies and concept hierarchies.

However, the extraction of hypernym-hyponym pairs from a corpus can be time-consuming, and reconstructing the hierarchical network of a domain can be an extremely complex process. A systematic semi-automatized approach would thus lighten the workload of terminographers and would lead to more efficient data processing. This paper presents the French EcoLexicon Semantic Sketch Grammar (ESSG-fr), a French hyponymic sketch grammar for Sketch Engine (SkE) (https://www.sketchengine.eu/) (Kilgarriff et al. 2004), which provides a user-friendly way of extracting hyponymic pairs in the form of word sketches (WSs).

modifiers of "tea"			verbs with "tea" as object		
afternoon	107	10.25	have	341	5.4
more	72	5.61	make	271	6.16
hot	61	8.17	drink	210	10.57

Figure 1: WSs of *tea* in the British National Corpus.

WSs are automatic corpus-derived summaries of a word's grammatical and collocational behavior (Kilgarriff et al., 2004). The default WSs provided by SkE represent different syntagmatic relations, such as verb-object, modifiers or prepositional phrases (Figure 1). However, the development of paradigmatic WSs is a timely contribution to the field of Terminology because they allow terminographers to perform a more efficient conceptual analysis of any corpus uploaded to SkE.

The remainder of this paper is structured as follows. Section 2 gives a brief description of the semantic relation of hyponymy and other issues related to terminology work and extraction methods, such as knowledge-rich contexts and knowledge patterns. Section 3 presents the first version of the ESSG-fr, after which, Section 4 discusses the evaluation of the WSs generated with it. Finally, Section 5 gives the conclusions that can be derived from this study.

2. Hyponymic knowledge patterns

2.1 Hyponymic relations

Hyponymy, also known as the *type_of* or *is_a* relation, can be defined as a relation of inclusion based on similarity. According to Lyons (1968), it is the most fundamental paradigmatic relation, and it includes both instantiation and subsumption relations. Hypernyms are broader than hyponyms, which means that the first includes the latter. For instance, APPLE is a type of FRUIT because the traits of FRUIT are included in APPLE. For a hypernym to exist, it must include at least two hyponyms (co-hyponyms).

Hypernymy typically gives rise to unilateral entailment whereby the hypernym entails the hyponym but not vice versa (Murphy and Koskela, 2010). Thus, in a text, a hyponym can generally be substituted by its hypernym and still convey the message, but not the other way around. This is due to transitivity, one of the main properties of hyponymy. Nevertheless, functional hyponymy and non-prototypical meanings can give rise to erroneous inferences (Cruse, 2002; Murphy, 2010).

According to Murphy (2003), hyponymy is central to many models of the lexicon for three reasons: (1) its inference-invoking nature; (2) its importance in definitions; (3) its relevance in selectional restrictions in grammars. In the same line, Barrière (2004) highlights the important role of hyponymy in categorization and property inheritance. Hyponymy can cause multiple inheritance in multidimensional concepts (León-Araúz, 2017; León-Araúz and San Martín, 2012). For example, since WATER can be regarded either as

a type of LIQUID or a type of MOLECULE., it is a concept that has different categorizations (Gil-Berrozpe, León-Araúz, and Faber, 2017, 2018), which means that its definition should be flexible (San Martín, 2016). Consequently, the extraction of hyponymic pairs is crucial for the design and population of many fields in a TKB. After all, hyponymy plays an important role in our conscious thinking about what a word means (Murphy, 2003).

2.2 Knowledge-rich contexts

One of the most common approaches to the efficient extraction of information from a corpus is to search for knowledge-rich contexts (KRCs). A KRC is a context indicating at least one item of domain knowledge that could be useful for conceptual analysis (Meyer, 2001). To find KRCs in corpora, knowledge patterns (KPs) are commonly used, since they are considered to be one of the most reliable methods for the extraction of semantic relations (Barrière, 2004; Bowker, 2003; Condamines, 2002; Lafourcade and Ramadier, 2016; Lefever, Kauter, and Hoste, 2014; Marshman, Morgan, and Meyer, 2002, *inter alia*).

KPs are the linguistic and paralinguistic patterns that convey a specific semantic relation (Meyer, 2001). For example, English KPs conveying hyponymy are x *is a kind of* y and x *and other* y, whereas French KPs include x *est un type de* y, x *et d'autres* y, etc. Terminographers use KPs in manual searches as seed words in combination with the terms under study. However, when KPs are formalized as grammars, more efficient queries can be performed.

KRCs are also a characteristic feature in TKBs, as users are generally provided with both usage and cognitive contexts. The formalization of KPs can thus lead to a better selection of KRCs, as not all of them are equally valuable (León-Araúz and Reimerink, 2019).

2.3 Knowledge-pattern-based word sketches

WSs are automatically generated when a corpus is compiled and annotated in SkE. They are based on sketch grammars codified in CQL (Corpus Query Language) (Jakubíček et al., 2010). CQL allows for the formalization of sketch grammar rules in the form of regular expressions combined with part-of-speech (POS) tags. Sketch grammars are thus a collection of CQL expressions that can be used to generate ready-made WSs. Since KPs are similar to regular grammar patterns, semantic WSs can be generated with the same logic. For instance, the English hyponymic KP *such as* followed and preceded by a noun can be formalized as follows: *1:"N.*" [tag!="V.*"]* [word="such"] [word="as"] [tag!="V.*"]* 2:"N.*"*, which means that a noun should be followed by any number of elements not being a verb, the words *such* and *as*, any number of elements not being a verb and another noun.

In León-Araúz, San Martín and Faber (2016), we developed 64 new sketch grammar rules focusing on the extraction of various semantic relations in English. They are grouped under the English EcoLexicon Semantic Sketch Grammar (ESSG-en) and are available in <http://ecolexicon.ugr.es/essg>. In a subsequent study (León-Araúz and San Martín, 2018), we evaluated their performance in the EcoLexicon English corpus of environmental science texts (León-Araúz, San Martín, and Reimerink, 2018). In this paper, we present a hyponymic sketch grammar for French corpora (ESSG-fr), as well as improved development and evaluation methods. Furthermore, this sketch grammar was developed with and tested in corpora belonging to different specialized domains. It is thus expected to be domain-independent.

3. Creation of the ESSG-fr

The methodology followed for the creation of the ESSG-fr was similar to that in León-Araúz and San Martín (2018) for ESSG-en. It has four main phases:

1. Collection: A list of possible KPs in natural language was drawn up from different sources.
2. Codification: The KPs were encoded as CQL rules, which subsequently enabled the SkE to generate the WSs.
3. Enrichment and refinement: Different variations were tested to improve the results.
4. Evaluation: An evaluation of the precision of the KPs allowed the retrieval of new KPs and modifications to be applied to the CQL rules.

To create the ESSG-fr, we added two new techniques. In the collection phase, a parallel corpus was used to find new KPs in French by querying the ESSG-en KPs. In addition, in the enrichment phase, the SkE thesaurus function was used to find new variants of the already encoded KPs.

3.1 Pattern collection

3.1.1 Patterns referenced by other authors

The starting point of our KP list was the KPs referenced by various authors (Auger, 1997; Aussenac-Gilles and Séguéla, 2000; Borillo, 1996; Lefeuvre, Coustot, Condamines, and Rebeyrolle, 2017; Rebeyrolle and Tanguy, 2000). We obtained 60 candidate KPs, which were divided into categories based on the ordering of the hyponym, hypernym and other elements (verbs, nouns, etc.) in the KP.

By grouping them, we were able to encode similar KPs simultaneously, thus dealing with the same difficulties at the same time. This also led us to even lump certain KPs in a single CQL rule when it was deemed appropriate. As will be seen below, this categorization is maintained in the final version of the ESSG-fr (Section 3.5).

3.1.2 Patterns found in a parallel corpus

As a new method for the discovery of KPs, we queried the 18 ESSG-en hyponymic KPs in an English-French parallel corpus (i.e. OPUS2) (Tiedemann, 2012). This was carried out in the following steps:

1. Advanced search in CQL was performed of each KP individually in the English corpus of OPUS2.
2. Sub-hits were hidden to avoid overlaps of various relations in the same sentence (i.e., duplicate phrases were filtered out).
3. The French concordances were consulted, which SkE had aligned with the English concordances in two columns.
4. Concordances were shuffled to assure that the lines came from different sources.

5. A manual analysis of the first 25 shuffled concordances was performed and the new KPs found were noted. Batches of 25 continued to be analyzed until no new KPs were found.

This technique allowed us to find 53 new candidate KPs in French. For instance, by querying the English KP "type of HYPER ranges from HYPO to HYPO" in CQL, we were able to detect the French KP "types d'HYPER vont de HYPO1 à HYPO2" (Table 1), which was further refined and enriched.

...the types of allegations ranged from inappropriate verbal conduct to sexual assault...
...les types d'allégation vont de propos inconvenants à l'agression sexuelle...

Table 1: Aligned concordances from OPUS2.

3.2 Pattern encoding

To create a preliminary CQL version of the grammar, the different iterations were tested on the multidomain French corpus created by Drouin (2010) for his Transdisciplinary Scientific Lexicon (TSL) project. The TSL corpus totals 4,373,546 words from PhD theses and scientific papers. It is divided in nine domain-specific subcorpora: Archaeology, Chemistry, Geography, History, Computer Science, Engineering, Law, Physics, and Psychology. This ensured the applicability of the resulting grammar to different specialized domains.

During the encoding phase and afterward, certain KPs collected in the previous step were split or lumped to better manage them. It is also important to note that any KP (e.g. "HYPER est un genre de HYPO", *"HYPER is a type of HYPO"*) can take many different forms in natural language. For instance, the verb "être" (*be*) can be in different tenses or it can be preceded by an auxiliary or modal verb; "genre" (*type*) may be modified by adjectives and adverbs. There might also be enumerations in the hypernym or hyponym position, and many other variations. All potential variations must be considered when developing grammar rules. To illustrate how we have addressed this issue, the definitive CQL representation of the KP "HYPO et d'autres HYPER" *("HYPO and other HYPER*) is reproduced and explained in Table 2. A sample of three concordances obtained with this CQL KP is also shown.

2:[tag="N.*" & lemma!=".*genre|.*sorte|.*espèce|.*variété|.*type|.*exemple|.*groupe|.*classe|.*catégorie|.*famille|.*mode|.*caste|.*division|.*race|.*collection|membre|nombre|un|deux|trois|quatre|cinq|six|sept|huit|neuf|dix|dix.*|onze|douze|treize|quatorze|quinze|seize|."][1] [tag="V.I.*|V.S.*|V.M.*|V.G.*|V.N.*|Fp.*|Fz|Fd.*|Fx.*"]{0,12}[2] [lemma="et|ou"][3] [tag!="V.I.*|V.S.*|V.M.*|V.G.*|V.N.*|Fp.*|Fz|Fd.*|Fx.*"]{0,5}[4] ([lemma="tout"]? [lemma="autre"]|[word="d'"][word="autres"])[5] ([lemma=".*genre|.*sorte|.*espèce|.*variété|.*type|.*exemple|.*groupe|.*classe|.*catégorie|.*famille|.*mode|.*caste|.*division|.*race|.*collection"][6] [tag="R.*" & lemma!="ne"]? "A.*|VMP.*"? [tag="R.*" & lemma!="ne"]?[7] ([word="et|ou"] [tag="R.*" & lemma!="ne"]? "A.*|VMP.*" [tag="R.*" & lemma!="ne"]?)? [lemma="de"])?[8] [tag="R.*" & lemma!="ne"]? "A.*|VMP.*"?[9] 1:[tag="NC.*" & lemma!=" .*genre|.*sorte|.*espèce|.*variété|.*type|.*exemple|.*groupe|.*classe|.*catégorie|.*famille|.*mode|.*caste|.*division|.*race|.*collection|membre|nombre|part"][10] & 1.lemma != 2.lemma[11]
[1] The hyponym is any noun other than "genre" (*type*), "sorte" (*sort*), etc. (or nouns ending in "genre", "sorte", etc., such as "sous-genre" (*subtype*)) and cannot be a single letter or any of the listed numbers. [2] Any element from 0 to 12 times that is not a verb form in indicative, subjunctive, imperative, gerund or infinitive, nor punctuation signs such as parentheses, colons, semicolons, etc. [3] "Et" (*and*) or "ou" (*or*). [4] Any element from 0 to 5 times that is not a verb in indicative, subjunctive, imperative, gerund or infinitive nor punctuation signs such as parentheses, colons, semicolons, etc. [5] Lemma "autre" (*other*) preceded by an optional lemma "tout" (*any*), or "d'autres" (plural *other*). [6] Any noun other than "genre", "sorte", etc. (or nouns ending in "genre", "sorte", etc.). [7] Optionally any adverb other than "ne" (*not*), followed by an optional adjective or participle, followed optionally by any adverb other than "ne". [8] "Et" or "ou" followed by an optional adverb other than "ne", followed by an adjective or participle, followed by an optional sequence of an optional adverb other than "ne" followed by lemma "de" (*of*), all of which is optional. [9] Optionally any adverb other than "ne" followed by an optional adjective or participle. [10] The hypernym is any common noun other than "genre", "sorte", etc. (or nouns ending in "genre", "sorte", etc.). [11] The hyponym and the hypernym cannot be the same lemma.
...utilisé pour la production d'ammoniac et de nombreux autres composés organiques... ... et autres décisions rendus par la Cour suprême et autres organes judiciaires... ...nette entre l'identification et d'autres processus proches comme l'incorporation...

Table 2: CQL grammar rule and its explanation.

3.3 Pattern enrichment and refining

The enrichment process consisted of testing each CQL rule with additional optional elements to detect new variations of the KP (e.g., an optional adjective in a position not previously accounted for).

We added a new method to this step. More specifically, new variants were found by using the option that allows the automatic inclusion of the results of the thesaurus within a CQL query. The SkE thesaurus retrieves words with similar WS results, which tend to be (near-) synonyms, antonyms, hypernyms, hyponyms and co-hyponyms (Rychlý, 2016). For instance, to find new lemmas that could fill the position of "principal" *(main)* in the KP "HYPO est le HYPER principal" ("*HYPO is the main HYPER"*), we queried the CQL KP including the thesaurus results for *principal (~"principal-j"* in CQL). This allowed us to find other productive variants of the KP, such as "HYPO est le HYPER majoritaire, optimal, idéal, parfait, etc.".

The refinement process consisted of detecting erroneous concordance lines obtained with the CQL rules, analyzing the source of the error, and applying the appropriate changes to the CQL rule.

3.4 Pattern evaluation

Pattern evaluation involved evaluating each CQL KP to determine if certain KPs needed to be further refined or discarded. Each KP was queried in the TSL corpus and, for each one, 20 random concordance lines were extracted (using SkE's *Get a random sample* feature) and evaluated. The number of correct concordances was used to estimate the precision of each KP. During the evaluation, KPs were modified if possible refinements were detected. In that case, the KP was re-evaluated after the modifications.

At the end of the first round of evaluation, KPs with a precision rate of less than 15% were again enriched and refined. Once enhanced, they were re-evaluated. Those with an accuracy rate of less than 10% were finally discarded. These included the following: "HYPO signifie HYPER"

(*"HYPO means HYPER"*) or "qualifier HYPO de HYPER" (*"consider HYPO to be a HYPER"*).
A low precision threshold (i.e. 10%) was set to prioritize recall. The reason for this was that users of the ESSG-fr will access the results organized in WSs, in which the potentially most relevant results are at the top of the list and the noisier ones at the bottom.

3.5 Definitive version of the patterns

The definitive KPs were included in the same grammar, which is available at <https://ecolexicon.ugr.es/essg/> along with instructions on how to apply it to any user-owned French corpus in SkE. Table 3 summarizes each of the 57 KPs, divided into 8 categories.

A: ...HYPO...EST...HYPER… A1) HYPO EST un GENRE de HYPER / A2) HYPO EST une SPÉCIALISATION de HYPER / A3) HYPO EST un HYPER qui / A4) HYPO EST un HYPER ADJ / A5) tout HYPO EST un HYPER / A6) HYPO EST (le plus|moins ADJ)|MEILLEUR (parmi|des) HYPER / A7) HYPO EST le HYPER (le plus|moins ADJ)|PRINCIPAL
B: ...VERB...HYPO...HYPER B1) APPELER HYPO DET HYPER / B2) DÉFINIR HYPO comme HYPER / B3) UTILISER HYPO comme|en tant que HYPER / B4) entendre par HYPO DET HYPER / B5) Sont DÉFINIS comme|parmi HYPO les HYPER / B6) Sont APPELÉS HYPO les HYPER
C: …HYPO…VERB…HYPER C1) HYPO (est) DÉFINI comme|parmi HYPER / C2) HYPO (est) utilisé|employé comme|en tant que HYPER / C3) HYPO se DÉFINIT comme|parmi HYPER / C4) HYPO s'utilise|s'emploie comme|en tant que HYPER / C5) HYPO qu'on DÉFINIT comme|parmi HYPER / C6) HYPO qu'on utilise|emploie comme|en tant que HYPER / C7) par HYPO on entend HYPER / C8) par HYPO il est entendu HYPER / C9) HYPO ENTRE dans le GENRE de HYPER
D: …HYPO…HYPER D1) le HYPO, HYPER / D2) HYPO et d'autres HYPER / D3) le HYPO, le (plus|moins ADJ)|MEILLEUR (des|parmi) HYPER / D4) le HYPO, le HYPER (le plus|moins ADJ)|PRINCIPAL / D5) DÉFINITION de HYPO comme HYPER / D6) UTILISATION de HYPO comme|en tant que HYPER
E: …HYPER…EST…HYPO E1) GENRE de HYPER est HYPO / E2) GENRE de HYPER qu'on APPELLE HYPO / E3) GENRE de HYPER s'APPELLE HYPO / E4) DET (plus|moins ADJ)|MEILLEUR (des|parmi) HYPER est HYPO / E5) DET HYPER (le plus|moins ADJ)|PRINCIPAL est HYPO / E6) DET MEILLEUR HYPER est HYPO
F) …VERB…HYPER…HYPO… F1) RECENSER NUM HYPER entre autres HYPO / F2) Font partie des HYPER les HYPO / F3) Sont INCLUS dans les HYPER les HYPO
G: ...HYPER...VERB...HYPO… G1) HYPER regroupe|rassemble les HYPO / G2) HYPER vont|varient de HYPO1 à HYPO2 / G3) GENRE de HYPER comprend|inclut HYPO / G4) GENRE de HYPER VERB sous|par le TERME HYPO
H: …HYPER…HYPO H1) HYPER, HYPO NOTAMMENT / H2) HYPER, NOTAMMENT HYPO / H3) HYPER Y COMPRIS HYPO / H4) HYPER c'est-à-dire|à savoir HYPO / H5) HYPER par exemple|tel que HYPO / H6) ACTION(=HYPER) de HYPO / H7) HYPER parmi lesquels HYPO / H8) HYPER depuis HYPO1 jusqu'à HYPO2 / H9) HYPER, tant HYPO1 que HYPO2 / H10) HYPER allant|variant de HYPO1 à HYPO2 / H11) PLUSIEURS HYPER dont|entre autres HYPO / H12) tout GENRE de HYPER, dont|entre autres HYPO / H13) DET plus|moins ADJ|MEILLEUR HYPER à savoir|c'est-à-dire DET HYPO / H14) DET HYPER (le plus|moins ADJ)|PRINCIPAL à savoir|c'est-à-dire DET HYPO / H15) parmi les HYPER, HYPO / H16) comme|en tant que HYPER, DET HYPO
ABBREVIATIONS ACTION = action; opération; propriété; effet; phénomène; processus; procédure; sentiment; péché; événement; évènement; rôle; situation; acte; valeur; problème; maladie; objectif; procédé; besoin; relation; réaction; nécessité; lien / ADJ = adjective/participle (tag) / APPELER = appeler; nommer; dénommer / CONSTITUER = constituer; consister; former; représenter / DÉFINIR = définir; percevoir; classer; catégoriser; identifier; interpréter; caractériser; présenter; considérer; reconnaître / DÉFINITION = définition; perception; classification; catégorisation; identification; interprétation; caractérisation; présentation; considération; reconnaissance / DET = determinant (tag) / ENTRER = entrer; rentrer; aller / ÊTRE = être; constituer; consister en; représenter / GENRE = (lemmas ending in) genre; sorte; espèce; variété; type; exemple; groupe; classe; catégorie; famille; mode; caste; division; race; collection / HYPER = hypernym (tag) / HYPO = hyponym (tag) / INCLURE = inclure; comprendre; classer; catégoriser / MEILLEUR = meilleur; pire; principal; seul; premier; vrai; unique; véritable; réel / NOTAMMENT = notamment; spécialement; surtout; particulièrement; spécifiquement; concrètement; précisément; justement; singulièrement; nommément; exactement; principalement; essentiellement; avant tout; en particulier; par exemple / NOTER = noter; préciser; indiquer; constater; mentionner; souligner; évoquer; rappeler; signaler; décrire; remarquer / NUM = numeral (tag) / PRINCIPAL = majoritaire; optimal; idéal; parfait; minoritaire; principal; prédominant; prépondérant; prioritaire; majeur; prééminent; primordial; supérieur; inférieur ; capital; crucial; dominant; central; primaire; primitif; original; essentiel; pertinent; clé; fondamental; favori; préféré; priorisé; visé / PLUSIEURS = plusieurs; différents; de nombreux; un certain nombre de; un grand nombre de; beaucoup de; divers; quelques / PRÉSENTER = présenter; trouver; montrer; retrouver / RECENSER = dénombrer; recenser; regrouper; mentionner; identifier; repérer; présenter; trouver; montrer; retrouver / REPRÉSENTER = désigner; représenter; symboliser; dénommer; définir; caractériser / SPÉCIALISATION = spécialisation; spécification; précision; détermination / TERME = terme; nom; appellation; dénomination; désignation; substantif; expression; vocable; mot / UTILISATION = utilisation; emploi; usage / UTILISER = utiliser; se servir de; employer / VERB = verb (tag) / Y COMPRIS = y compris; incluant; tout en comptant; sauf; hormis; exception faite de; à l'exception de; excepté

Table 3: Summary of the 57 ESSG-fr KPs.

When the ESSG-fr is applied to a corpus and the user queries a term for WSs, they obtain two WSs generated from the ESSG-fr: X *est le générique de…* (*X has subtype…*), which provides a list of candidate hyponyms, and X *est un type de…* (*X is a type of…*), which provides a list of candidate hypernyms (Figure 2). Table 4 shows a sample of the concordances associated with the WS in Figure 2 (accessible to the user by clicking on the frequency number (in light blue).

"composé" est le générique de...				"composé" est un type de...			
acide	23	10.62	•••	solvant	3	8.55	•••
eau	7	9.15	•••	réaction	3	6.17	•••
carbone	4	8.58	•••	formule	2	10.19	•••

Figure 2: ESSG-fr WSs for "compose" (*compound*) in a Chemistry corpus.

...Bronsted et Thomas Lowry, un acide est un composé chimique qui tend à donner...
...Certains composés tels que les lactones et les acides gras qui ont des points...
... chimie générale au cas de l'oxygène et de ses composés, notamment de l'eau...
...L'eau à l'état liquide est un composé amphotère à la fois acide et base au sens de...
...mises en évidence dans des usines où le composé était utilisé comme solvant...

Table 4: Concordances associated to the WS for "composé".

4. Word sketch evaluation

The ESSG-fr's performance was assessed by evaluating its resulting WSs, which are the main product that users will obtain from it.

4.1 Corpus compilation

To avoid evaluating the grammar with the same corpus with which it was trained, we compiled three new domain-specific corpora (Chemistry, Law and Psychology) using the SkE corpus creation tool from keywords. This tool automatically finds and downloads texts from the web. The keywords used for the corpus compilation were the same ones later used to evaluate the WSs. This ensured that the evaluated terms were contained in sufficient frequency in the corresponding corpus.

The list of terms for each domain is reproduced in the first column of Tables 6, 7 and 8. To select the terms, the three TSL subcorpora corresponding to each domain (Chemistry, Law and Psychology) were fed to the term extractor Termostat (http://termostat.ling.umontreal.ca/) (Drouin, 2003) to obtain the nouns with the highest specificity score. For each corpus, we retained 20 terms. To ensure termhood and domain specificity, each retained term needed to be included in Termium Plus (https://btb.termiumplus.gc.ca/) (official termbank of the Canadian Government) accompanied by a definition and labeled as pertaining to the corresponding domain.

To compile the corpora, SkE default parameters were used except for *Max URLs per search*, which was set to 30 instead of 20 to obtain larger corpora. After automatically removing duplicate sentences, the Chemistry corpus had 3,037,251 words; the Law corpus, 2,636,444 words, and the Psychology corpus, 6,662,299 words. Despite the considerable difference in size, the decision was made not to intervene so that each corpus was automatically created by the same parameters.

4.2 Evaluation methodology

Once each corpus was compiled using the ESSG-fr, the 20 terms were queried with the WS feature. For each term, we evaluated the three most frequent results in either the *générique* WS or the *type* WS. Since certain terms were more likely to return more relevant results concerning their hypernyms or hyponyms (i.e. depending on their conceptual granularity), for each term, we only evaluated the WS with the first three most frequent results. For example, the psychological term "symptôme" (*symptom*) returned the results in Figure 3. The first three results of the hyponym list (i.e. the *générique* WS) add up to 20, whereas the hypernym list (i.e. the *type* WS) adds up to 12. Therefore, we evaluated its hyponyms. The logic of this approach was that the WS with more results was more likely to contain more useful information for the user. For instance, in this case, a user will be more likely to consult which types of "symptôme" there are (*"symptôme" est le générique de...*), than its hypernyms (*"symptôme" est un type de...*), since "symptôme" is a top-level concept whose hypernyms would normally be of little or no interest.

"symptôme" est le générique de...				"symptôme" est un type de...			
hallucination	8	9.53	•••	jouissance	4	9.07	•••
état	6	8.92	•••	formation	4	8.9	•••
pensée	6	8.62	•••	sens	4	8.59	•••

Figure 3 : ESSG-fr WSs for "symptôme" in a Psychology corpus.

In order to determine the precision of the WSs, we assessed whether the results were correct by accessing the corresponding concordance lines. A concordance line is considered correct (i.e. a true positive) if a hyponymic relation between the two terms is explicitly expressed.

It often occurred that even though the identified relation was correct, it had been mistakenly extracted. In those cases, the concordance was evaluated as incorrect. For example, in Table 5, the first concordance (from the Psychology corpus) is a false positive because the strict correct relation would be "pulsion" (*drive*) is a type of "source", even though the relation "pulsion" is a type of "énergie" (*energy*) is indeed a valid relation, as shown in the second concordance.

...dans la névrose, la pulsion sexuelle est la source d'énergie la plus importante et la...
... de poussée. La pulsion est une énergie, une force motrice. Mais cette énergie...

Table 5: Concordances for "*pulsion* is a type of *énérgie*".

Many of the false positives were caused by the inherent limitations of KP-based extraction of semantic relations with WSs, such as POS-tagger mistakes, polysemy, anaphora, etc. (see León and San Martín (2018) for examples).

It should be noted that some results only make sense by accessing the concordances as they are part of a nominal compound. For example, "eau" (*water*) is the third extracted hyponym of "solution" in Chemistry. However, in all of the four associated concordances, it is part of either "eau de Javel" (*bleach*) or "eau oxigénée" (*hydrogen peroxide*). Both are indeed types of solution and were assessed as true positives.

4.3 Results

Tables 6, 7 and 8 show the results of the evaluation of the WS obtained with each corpus. In each line, the search term appears followed by G or T, depending on whether the *générique* or *type* WS was evaluated. In the following columns, the first three results with their frequency and precision rate are reproduced. In total, 180 pairs of terms were evaluated.

As can be observed in Figure 4, the total average precision of the first three results per term is 61.07%. In Psychology, the precision rate was 78.34%; in Chemistry 68.46%; and in Law 36.44%. The difference between the domains probably has two origins, the first being the size of the corpus, since there is a clear correlation between size and precision. It is also possible to hypothesize that the peculiarities of legal language itself may make it necessary to create domain-specific KPs to improve precision.

Search term	WS	1st result	F	P%	2nd result	F	P%	3rd result	F	P%
molécule	G	acide	13	84.7	hydrogène	6	0	pyridine	5	80
adsorption	T	phénomène	41	95.1	processus	36	100	propriété	9	100
composé	G	acide	23	86.9	eau	7	57.1	carbone	4	0
complexe	T	édifice	5	100	catalyseur	4	100	donneur	3	100
ligand	T	rôle	14	100	carbène	2	100	fragment	2	50
atome	G	ion	8	100	hydrogène	6	50	carbone	4	50
éther	T	solvant	8	75	réaction	2	0	sou	1	0
ion	T	atome	8	75	solution	7	0	potassium	3	0
oxygène	T	élément	5	80	atome	3	66.6	agent	2	100
concentration	T	paramètre	11	100	valeur	11	90.9	facteur	9	100
phase	T	solvant	6	50	gaz	5	60	liquide	4	100
réaction	G	réduction	62	98.4	substitution	49	100	rétro-aldolisation	43	100
cation	G	ion	3	100	calcium	2	100	sodium	2	100
acide	T	composé	23	82.6	molécule	13	84.6	solvant	8	87.5
solution	G	ion	7	0	électrolyte	4	100	eau	4	50
solvant	G	eau	19	94.7	éther	8	75	acide	8	87.5
argon	T	gaz	5	60	prénom	1	0	extinction	1	0
lixiviation	T	processus	23	95.6	procédé	10	90	phénomène	6	100
électron	T	orbitale	3	0	particule	3	100	état	2	100
éthylène	T	molécule	4	50	propriété	3	0	bilan	2	0

Table 6: Evaluation of the WSs obtained with the Chemistry corpus.

Search term	WS	1st result	F	P%	2nd result	F	P%	3rd result	F	P%
article	T	disposition	6	83.3	préfix	4	0	droit	3	0
compétence	T	problème	6	100	juridiction	4	0	autorité	3	0
pouvoir	T	désequilibre	2	0	tâche	2	0	décision	2	0
droit	T	acte	14	0	norme	9	66.7	effet	7	0
organe	G	conseil	8	100	cour	6	100	aide	3	0
disposition	T	acte	23	0	règlement	4	0	principe	3	0
jurisprudence	T	source	6	66.7	droit	3	0	ombre	1	100
juge	T	rôle	8	100	gravité	4	0	magistrat	4	50
constitution	T	norme	17	76.5	texte	7	100	procédure	5	60
fondement	G	démocratie	2	0	constitution	2	100	territorialité	1	100
souveraineté	T	acte	6	0	idée	2	100	qualité	2	100
alinéa	T	portion	1	100	web	1	0	mine	1	0
juridiction	G	cour	9	88.9	tribunal	7	100	compétence	4	0
conception	T	technologie	2	0	partialité	1	0	origine	1	0
règlement	T	procédure	35	94.3	processus	7	100	acte	5	100
clause	T	contrat	7	0	disposition	6	100	irrégularité	3	0
litige	T	primauté	1	0	incompétence	1	0	saut	1	0
constitutionnalité	T	contrôle	4	0	juge	2	0	détermination	1	0
doctrine	T	entité	2	0	notion	2	0	not	1	0
législation	G	excusabilité	1	0	risque	1	0	crédit	1	100

Table 7: Evaluation of the WSs obtained with the Law corpus.

Search term	WS	1st result	F	P%	2nd result	F	P%	3rd result	F	P%
moi	T	instance	4	100	objet	4	50	identité	2	100
pulsion	T	concept	8	100	énérgie	4	75	principe	4	75
sujet	G	enfant	9	66.7	personne	5	80	signifiant	4	100
psychopathologie	G	dépression	6	100	mmpi	1	0	compléxité	1	0
affect	G	angoisse	8	75	honte	4	100	douleur	3	100
dépression	T	maladie	24	95.8	trouble	13	69.2	problème	9	100
dépendance	T	situation	65	100	relation	42	100	lien	17	94.1
identification	T	processus	15	93.3	mécanisme	11	90.9	objet	10	0
symptôme	G	hallucination	8	75	état	6	16.7	pensée	6	16.7
surmoi	T	instance	9	100	héritier	3	100	conscience	2	50
représentation	T	processus	15	86.7	forme	6	100	objet	4	25
socialisation	T	processus	69	100	phénomène	7	100	procédure	2	100
élaboration	T	processus	5	100	image	2	0	théorie	2	0
schizophrénie	T	maladie	33	97	psychose	14	92.9	trouble	12	75
refoulement	T	processus	31	100	mécanisme	13	100	défense	8	25
centralité	T	valeur	3	0	sentiment	3	100	qualité	2	100
besoin	G	sécurité	46	97.8	répit	34	94.1	compensation	28	100
ambivalence	T	sentiment	3	100	mot	2	100	mécanisme	2	100
encodage	T	processus	7	100	opération	5	100	assurance	1	100
introjection	T	processus	15	100	mécanisme	6	83.3	concept	3	100

Table 8: Evaluation of the WSs obtained with the Psychology corpus.

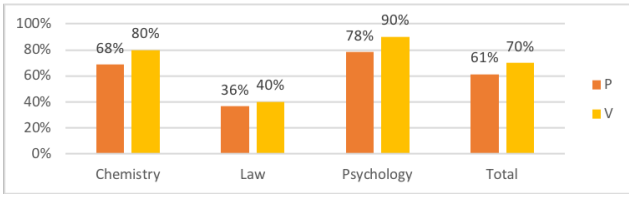

Figure 4: Average precision (P) and validity (V).

We also calculated a parameter that we have called *validity*, according to which a result is valid when at least one of its associated concordances is a true positive. For example, the relation "pensée" (*thinking*) is a type of "symptôme" is considered valid although only one of its concordances is a true positive (the first one in Table 9):

...que des variations du symptôme fondamental obsessionnel, à savoir la pensée...
...diffusion de la pensée était le symptôme le plus fréquent (37 %) alors que les...
...symptômes existants de dépression et spécialement des styles de pensée négative...
les symptômes positifs tels que les hallucinations, le délire, les troubles de la pensée...
...partage des pensées est un symptôme apparenté, dans lequel le patient a...
... symptômes de désorganisation, tels que les troubles du cours de la pensée...

Table 9: Concordances for "*pensée* is a type of *symptôme*"

The average validity rate is 70% across domains and is as high as 90% in the case of Psychology. Furthermore, 100% of terms in Chemistry and Psychology have at least one valid result among their three results, and 70% of terms in Law have at least one valid result.

In regard to the precision and validity by order of result, the first result (i.e. the one with the highest number of concordances) is the one that provides the highest precision (68.75%) (Figure 5) and validity (76.67%) (Figure 6) in total. This trend is also present across domains. In all cases, the second and third results score slightly worse than the first one, and in most cases, the second one scores slightly better than the third one.

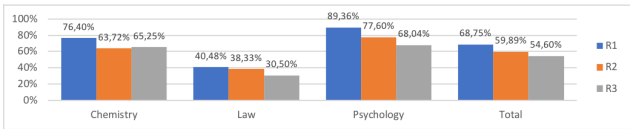

Figure 5: Precision of first, second and third result.

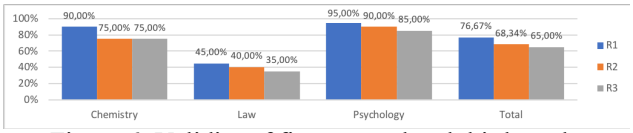

Figure 6: Validity of first, second and third result.

In 73.33% of the cases, the column analyzed was the hypernymic one ("est un type de" WS) (T in Tables 6-8). This predominance of the hypernyms was also observed in all domains (Chemistry: 65%, Law: 80%; Psychology: 75%). However, further studies will be needed to validate this trend and, if confirmed, to explain its causes. In terms of precision and validity, it is, however, the hyponym list ("est le générique de" WS) (G in in Tables 6-8) that shows

greater precision and validity (Figure 7). However, this only occurs in Chemistry and in Law, since in Psychology the opposite is observed.

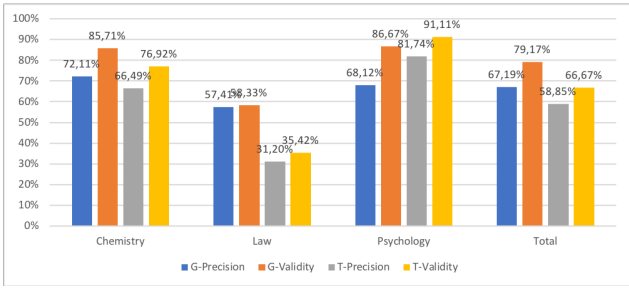

Figure 7: Precision and validity of the hypernyms (T) and hyponyms (G) lists.

If we analyze the differences of precision and validity according to the order of results between the list of hyponyms (G) and hypernyms (T), the precision in the list of hypernyms decreases more with the second and third results (-9.93% and -15.12%, respectively) compared to the first result than in the list of hyponyms (-5.94% and -11.49%, respectively) (Figure 8). This difference is even more pronounced in the case of validity: -11.36% and -13.63% for the hypernym list; -0 % and -6.25% for the hyponym list. This indicates that in the terms where hypernyms predominate, a single result tends to stand out in precision and validity. In contrast, in terms where hyponyms predominate, there is a wider range of possibilities. This is consistent with the fact that, for a hyponymic relation to exist, it is sufficient for a term to have one hypernym (and one co-hyponym); but it must have at least two hyponyms.

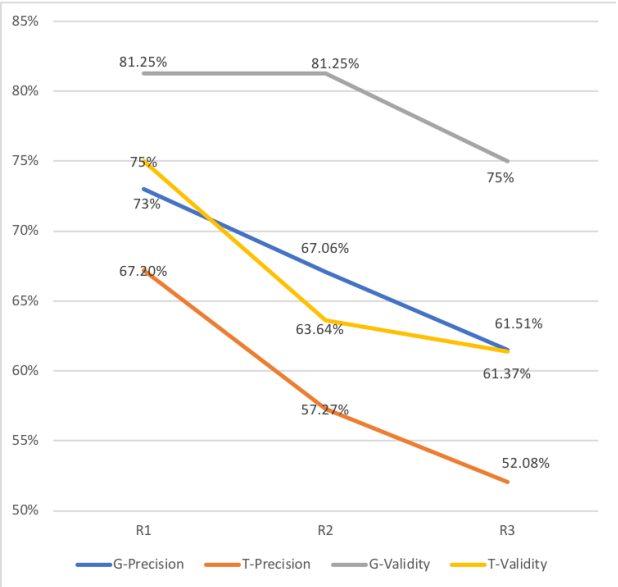

Figure 8: Precision and validity according to the order of results in the list of hyponyms (G) and hypernyms (T).

5. Conclusions and future work

This paper has presented the development and evaluation of the ESSG-f. The results of the evaluation show that in its current state the ESSG-fr can be very useful for terminographers and other users who may need to extract hyponymic relationships from specialized corpus in French. However, the lower precision in the legal corpus indicates the need to analyze whether legal language requires specific patterns or whether the results are due to the smaller size of the corpus.

The ESSG-fr is also a useful tool for analyzing the workings of the hyponymy relation itself, particularly the differences in the detection of hypernyms and hyponyms in corpora. Likewise, the ESSG-fr allows the study of semantic phenomena such as multidimensionality.

In future work, we will continue to improve the precision and recall of the ESSG-fr. In particular, we will focus on the study of the prevalence and precision of the different KPs depending on the domain. We also plan to explore the possibility of creating domain-specific versions of the ESSG-fr.

Furthermore, future studies will analyze which type of terms offers greater precision in the extraction of their possible hypernyms or hyponyms and their causes (i.e. granularity or abstraction levels, processes versus objects, etc.). The ESSG-fr will also be enriched with other types of semantic relationships such as meronymy and cause.

Finally, the methodological innovations used in this paper will be applied to improve the ESSG-en. More specifically, the KPs of the French grammar will be used to extract new English KPs from a parallel corpus.

6. Acknowledgements

We thank Patrick Drouin for allowing us to use his TSL French corpus.

This research was carried out as part of project *Développement d'une méthodologie d'élaboration de définitions terminologiques : analyse de corpus et variation contextuelle* (2020-NP-267503) funded by Quebec's Society and Culture Research Fund (FQRSC) and *Herramientas terminológicas orientadas hacia la traducción de textos medioambientales* (FF2017–52740-P) funded by the Spanish Ministry of Economy and Competitiveness.

7. References

Auger, A. (1997). *Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles*. PhD Thesis. University of Neuchâtel

Aussenac-Gilles, N., and Séguéla, P. (2000). Les relations sémantiques : du linguistique au formel. *Cahiers de Grammaire*, 25, 175–198.

Barrière, C. (2004). Knowledge-Rich Contexts Discovery. *Seventeenth Canadian Conference on Artificial Intelligence*, 187–201. London, Ontario: CSCSI.

Borillo, A. (1996). Exploration automatisée de textes de spécialité : repérage et identification de la relation lexicale d'hyperonymie. *Linx*, (34–35), 113–124.

Bowker, L. (2003). Lexical Knowledge Patterns, Semantic Relations, and Language Varieties: Exploring the Possibilities for Refining Information Retrieval in an International Context. *Cataloging & Classification Quarterly*, 37(1–2), 153–171.

Condamines, A. (2002). Corpus analysis and conceptual relation patterns. *Terminology*, 8(1), 141–162.

Cruse, A. (2002). Hyponymy and its varieties. In R. Green, C. A. Bean, and S.H. Myaeng (Eds.), *The Semantics of Relationships: An Interdisciplinary Perspective* (pp. 3–22). Dordrech: Kluwer Academic Publishers.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1), 99–115.

Drouin, P. (2010). Extracting a bilingual transdisciplinary scientific lexicon. In S. Granger and M. Paquot (Eds.), *eLexicography in the 21st century: new challenges, new applications* (pp. 43–53). Louvain-la-Neuve: Presses Universitaires de Louvain.

Gil-Berrozpe, J. C., León-Araúz, P., and Faber, P. (2017). Specifying Hyponymy Subtypes and Knowledge Patterns: A Corpus-based Study. In I. Kosem, J. et al. (Eds.), *eLexicography in the 21st century. Proceedings of eLex 2017 conference* (pp. 63–92). Brno: Lexical Computing.

Gil-Berrozpe, J. C., León-Araúz, P., and Faber, P. (2018). Subtypes of Hyponymy in the Environmental Domain: Entities and Processes. In C. Roche (Ed.), *Proceedings of TOTh 2016* (pp. 39–54). Chambéry: Université Savoie-Mont-Blanc.

Jakubíček, M., Kilgariff, A., McCarthy, D., and Rychlý, P. (2010). Fast syntactic searching in very large corpora for many languages. *Proceedings of PACLIC 24*, 741–747.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The Sketch Engine. In G. Williams and S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 105–116). Lorient: EURALEX.

Lafourcade, M., and Ramadier, L. (2016). Semantic Relation Extraction with Semantic Patterns Experiment on Radiology Reports. In N. Calzolari et al. (Eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 4578–4582). Portorož, Slovenia: ELRA.

Lefeuvre, L., Coustot, K., Condamines, A., and Rebeyrolle, J. (2017). *MAR-REL : Liste de candidats-marqueurs français pour les relations d'hyperonymie, de méronymie et de cause*. http://redac.univ-tlse2.fr/misc/mar-rel/Liste-des-marqueurs.pdf

Lefever, E., Kauter, M. Van de, and Hoste, V. (2014). HypoTerm: Detection of hypernym relations between domain-specific terms in Dutch and English. *Terminology*, 20(2), 250–278.

León-Araúz, P. (2017). Term and concept variation in specialized knowledge dynamics. In P. Drouin, A. et al. (Eds.), *Multiple perspectives on Terminological Variation* (pp. 213–258).

León-Araúz, P., and Reimerink, A. (2019). High-density knowledge rich contexts. *Argentinian Journal of Applied Linguistics*, 7(1), 109–130.

León-Araúz, P., and San Martín, A. (2012). Multidimensional Categorization in Terminological Definitions. In R. V. Fjeld and J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 578–584). Oslo: EURALEX.

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

León-Araúz, P., San Martín, A., and Faber, P. (2016). Pattern-based Word Sketches for the Extraction of Semantic Relations. *Proceedings of the 5th International Workshop on Computational Terminology*, 73–82. Osaka.

León-Araúz, P., San Martín, A., and Reimerink, A. (2018). The EcoLexicon English Corpus as an open corpus in Sketch Engine. In J. Čibej et al. (Eds.), *Proceedings of the 18th EURALEX International Congress* (pp. 893–901). Ljubljana: Euralex.

Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

Marshman, E., Morgan, T., and Meyer, I. (2002). French patterns for expressing concept relations. *Terminology*, 8(1), 1–29.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. In D. Bourigault, C. Jacquemin, and M.C. L'Homme (Eds.), *Recent advances in computational terminology* (pp. 279–302). Amsterdam: John Benjamins.

Murphy, M. L. (2003). *Semantic Relations and the Lexicon*. Cambridge: Cambridge University Press.

Murphy, M. L. (2010). *Lexical Meaning*. Cambridge: Cambridge University Press.

Murphy, M. L., and Koskela, A. (2010*). Key terms in Semantics*. London, New York: Continuum.

Rebeyrolle, J., and Tanguy, L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de grammaire*, 25, 153–174.

Rychlý, P. (2016). Evaluation of the Sketch Engine Thesaurus on Analogy Queries. In A. Horák, P. Rychlý, and A. Rambousek (Eds.), *Tenth Workshop on Recent Advances in Slavonic Natural Language Processing*, RASLAN 2016 (pp. 147–152).

San Martín, A. (2016). *La representación de la variación contextual mediante definiciones terminológicas flexibles*. PhD thesis. University of Granada.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 2214–2218.