# Automatic Lexical Collocate Extraction for Corpus-Based Ontology Building and Refinement: A FunGramKB Case Study of the THEFT Conceptual Scenario

Nicolás José Fernández-Martínez and Ángel Miguel Felices-Lago
Universidad Católica San Antonio de Murcia / University of Granada

Traditional corpus-based methods rely on manual inspection and extraction of lexical collocates in the study of selection preferences, which is a very costly, labor-intensive, and time-consuming task. Devising automatic methods for lexical collocate extraction becomes necessary to handle this task and the immensity of corpora available. With a view to leveraging the *Sketch Engine* platform and in-built corpora, we propose a working prototype of a Lexical Collocate Extractor (LeCoExt) command-line tool that mines lexical collocates from all types of verbs according to their syntactic constituents and Collocate Frequency Score (CFS). This might be the first tool that performs comprehensive corpus-based studies of the selection preferences of individual or groups of verbs exploiting the capabilities offered by *Sketch Engine*. This tool might facilitate the task of extracting rich lexico-semantic knowledge from diverse corpora in a few seconds and at a click away. We test its performance for ontology building and refinement departing from a previous detailed analysis of stealing verbs carried out by Fernández-Martínez & Faber (2020). We show how the proposed tool is used to extract conceptual-cognitive knowledge from the THEFT scenario and implement it into FunGramKB Core Ontology through the creation and modification of theft-related conceptual units.

**Keywords:** lexical collocate, selection preferences, automatic extraction, corpus-based, *Sketch Engine,* FunGramKB, stealing verbs, theft scenario, ontology building

## 1. Introduction

The aim of this paper is two-fold. First, we propose the creation of the Lexical Collocate Extractor (LeCoExt)[1], a working prototype of a command-line tool for automatic corpus-based extraction of lexical collocates from *Sketch Engine* corpora. Then, we apply this tool to mine lexical collocates of stealing verbs and build rich semantic knowledge for the implementation and refinement of the conceptual scenario of THEFT into the FunGramKB Core Ontology (Periñán-Pascual & Arcas-Túnez, 2010a, 2010b; Periñán-Pascual, 2013)[2]. We depart from a comprehensive analysis of the lexico-semantic dimensions of stealing verbs (Fernández-Martínez & Faber, 2020) and previous work on theft and crime conceptualization in the FunGramKB Satellite Ontology (Felices-Lago, 2015, 2016). With this automatic extraction method, we explore new horizons in corpus-based approaches to ontology building that demand improved methodological tools and guidelines (Jiménez-Briones & Luzondo-Oyón, 2011). In this respect, we set up the foundations of our corpus-based tool on computational methods for a comprehensive study of selectional preferences. We then leverage our corpus findings to build rich semantico-conceptual knowledge and, at the same time, model deep-semantic conceptual relations and specifications of the THEFT scenario, using for that purpose the in-built FunGramKB Conceptual Representational Language known as COREL (Periñán-Pascual & Mairal-Usón, 2010). This results in the creation of new #EVENT concepts[3], such as $SHOPLIFT_00, $RUSTLE_00 and $PLUNDER_00, and the modification of the extant concepts +STEAL_00 and $EMBEZZLE_00 in the FunGramKB Core Ontology to account for the whole spectrum of stealing acts.

---

[1] The original source code as well as a download link for the application can be found on https://github.com/njfm0001/LeCoExt/

[2] FunGramKB is a multipurpose lexico-conceptual knowledge base for NLP systems whose characteristics will be explained in section 2.2.

[3] #EVENT is a metaconcept that stands for all verbs in FunGramKB.

The remainder of this paper is organized as follows. Section 2 introduces the lexico-semantic dimensions of stealing verbs, FunGramKB and its Ontology. Section 3 describes the proposed automatic extraction method and a step-by-step explanation of the LeCoExt tool applied to stealing verbs. Section 4 provides the results obtained from our automatic corpus search and discusses their implementation into the FunGramKB Core Ontology. Some conclusive remarks derived from this research are offered in section 5.

## 2. Theoretical background

### 2.1. Stealing verbs: a lexico-semantic perspective

Stealing verbs refer to the illicit activity carried out by a thief that illegally takes something that belongs to someone or from a place (Boas, 2013). The THEFT domain is a subdomain embedded in the POSSESSION domain (Faber & Mairal-Usón, 1999), and covers a great range of verbs, such as *steal, rob, pilfer, pinch, embezzle, filch, plunder, loot*, etc.

Stealing verbs share a common conceptual space determined by the theft conceptual scenario. In Frame Semantics (Fillmore & Baker, 2010; Ruppenhofer et al., 2017), the *Theft* and *Robbery* frames encapsulate stealing verbs in terms of the prototypical cognitive-conceptual patterns of theft events: their participants or 'perpetrators', the things stolen or 'goods' and the human and/or locational origin of the stealing act, also called 'victim' or 'source'. These are the core frame elements of stealing verbs, which can also be described as semantic roles. All these verbs share a common 'background frame', while at the same time they specify their own lexico-semantic idiosyncrasies derived from their encyclopedic knowledge (Goldberg, 2010). This means that the underlying semantics of stealing verbs specify which aspects or participants are more prominent than others. For instance, *steal* puts greater emphasis on the perpetrator and the goods, and *rob* on the perpetrator and the victim or source (Dux, 2018; Thorgren, 2005). Their profiling character is ultimately reflected in the syntactic behavior of their arguments, which means that the profiled semantic roles typically correspond to compulsory syntactic elements in the clause, whereas the non-profiled elements occupy a less important syntactic role. Thus, in *steal*, the perpetrator and the goods correspond to the subject and direct object, its compulsory syntactic elements, whereas the victim or source corresponds to an optional oblique object introduced by the preposition *from*. In *rob*, the perpetrator and victim or source correspond to the subject and direct object respectively whereas the goods are usually left unspecified, though they may surface as an oblique object introduced by the preposition *of*.

As for the syntax of stealing verbs, most are monotransitive with optional oblique objects (Berman, 1982). There is a subgroup of stealing verbs called plundering verbs that exhibit complementary syntactic patterns: *plunder* (*plunder something from someone/a place* or *plunder a place*); *pillage* (*pillage something from someone/a place* or *pillage a place*); *loot* (*loot something from someone/a place or loot a place*); and *despoil* (*despoil someone/a place of something* or *despoil something*). Each of these patterns either focus on the victim or source, or on the goods.

Another important aspect that defines stealing verbs is their selection preferences. In a list of the most frequent lexical collocates that go with verbs of stealing, special attention should be paid to those collocates that appear as syntactic arguments of such verbs, that is, as subjects, direct objects, and oblique objects. These syntactic arguments must be in turn associated to their semantic roles as perpetrator, goods, victim, and source. On the basis of such lexico-semantic knowledge, we can then provide a conceptual characterization of the THEFT scenario and the commonest conceptual categories associated to each semantic role.

### **2.2.** FunGramKB: definition, scope and architecture

The web is flooded with vast volumes of data, mostly natural language data, which can be exploited for a great range of applications. Intelligent systems based on Natural Language Processing (NLP) and Artificial Intelligence (AI) address this massive amount of data and, though performing relatively well, still miss out on the rich semantico-conceptual knowledge that underlies natural languages. A representational model based on deep semantics could help machines and algorithms to understand the intricacies of human languages, in a way that mimics how humans process and react to language stimuli. FunGramKB arose out of these concerns in the early work of Periñán-Pascual & Arcas-

Túnez (2004, 2005), with the aim of providing machines with a linguistic-based reasoning engine endowed with deep-semantic knowledge to decipher the nature of language beyond traditional AI probabilistic-based algorithms. FunGramKB has become "a multipurpose lexico-conceptual knowledge base for NLP systems" (Periñán-Pascual & Mairal-Usón, 2009, p. 266; Periñán-Pascual & Arcas-Túnez, 2010a; Periñán-Pascual, 2012) that intends to cover a wide range of NLP applications, such as text mining, conversational agents, dialogue-based systems, machine translation, etc. A knowledge base is defined as a computational repository of conceptual and linguistic units that captures rich lexical, semantic, and morphosyntactic information. FunGramKB adopts a representational model based on deep semantics (Velardi et al., 1991). Unlike traditional surface-semantic approaches such as *WordNet* (Miller & Fellbaum, 2007), FunGramKB provides, not only semantic relations among conceptual units, but also conceptual definitions about the semantic types and arguments that take part in the conceptual-cognitive patterns derived from our encyclopedic knowledge.

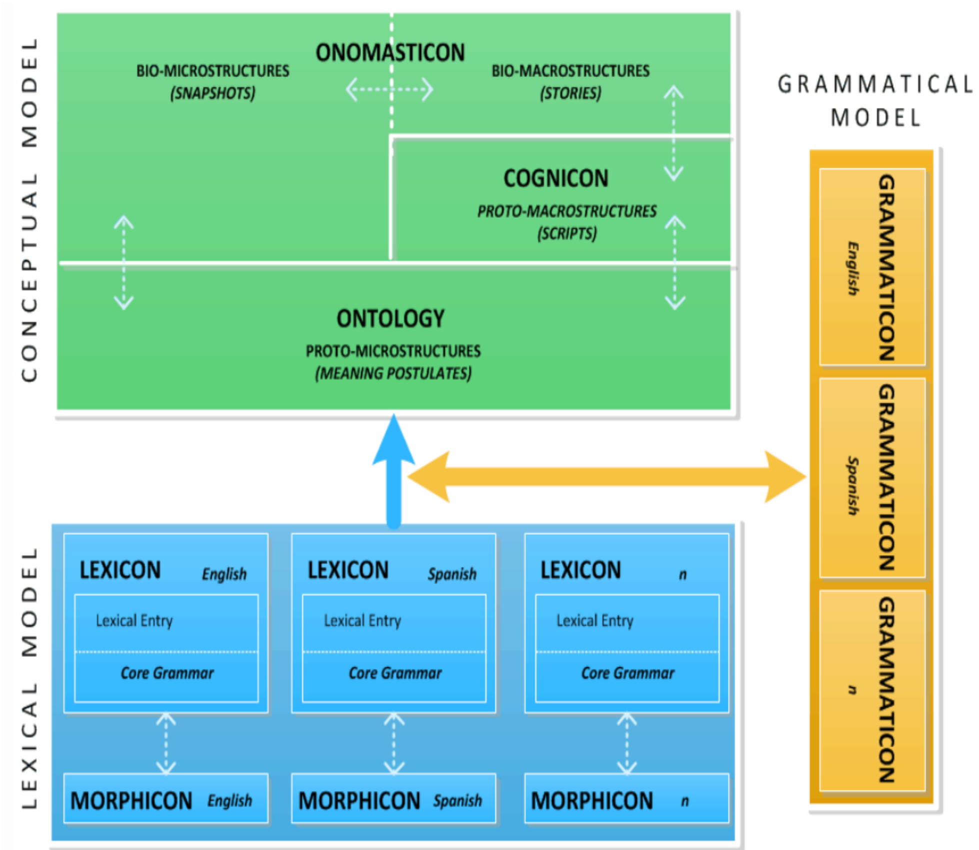



**Figure 1.** The FunGramKB modular architecture

Figure 1 presents the FunGramKB architecture, made of three main modules. The lexical model handles the different lexica, each with its language-specific morphosyntactic and pragmatic information about lexical units. The morphicon deals with cases of inflectional morphology. At the grammatical level, the grammaticon, based on Role and Reference Grammar (Van Valin, 2005) and the Lexical Constructionist Model principles (Ruiz-de-Mendoza Ibáñez & Mairal-Usón, 2009), handles the language-specific constructional aspects and the semantics-syntax mapping algorithm. The conceptual level stores non-linguistic knowledge through the cognicon, which contains procedural knowledge in the form of scripts that exemplify sequences of stereotypical life situations (e.g., 'buying at a supermarket', 'going to work'); the onomasticon, which incorporates real-world knowledge about instances and events through snapshopts and stories (e.g., *Notre Dame fire*, *president Obama*, *the Alhambra*); and the ontology, which provides universal concepts and their prototypical conceptual-cognitive patterns derived from the semantic knowledge embedded in our minds.

### **2.3.** The FunGramKB ontology

The ontology, made of a Core Ontology and satellite or domain-specific ontologies, provides a hierarchical arrangement of conceptual units into four levels, organized in a stepwise conceptual fashion, using the symbols #, +, $, and -, and assigning a numerical code to these levels (00, 01, 02, etc.), except to subconcepts:

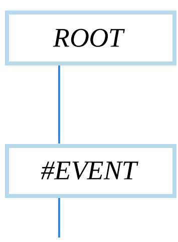
a. *Metaconcepts*, preceded by the symbol # (e.g., #EVENT, #ENTITY, and #QUALITY., etc.), constitute the upper level in the taxonomy and coincide with some of the most relevant linguistic ontologies, e.g., SIMPLE (Pedersen & Keson 1999, Lenci et al., 2000), SUMO (Niles & Pease, 2001), DOLCE (Gangemi et al., 2005; Masolo et al., 2003), among others. The FunGramKB Ontology is divided into three subontologies, where subsumption (IS-A) is the only taxonomic relation permitted. Each of these subontologies arranges lexical units in terms of different parts of speech: #ENTITIES for nouns (e.g., +SIN_00, +SOUL_00, +AIRCRAFT_00, etc.), #EVENTS for verbs (e.g., +BUILD_00, +THINK_00, etc.), and #QUALITIES for adjectives and some adverbs (e.g., +DIRTY_00,

+ALONE_00, etc.).

b. *Basic concepts*, preceded by symbol + (e.g., +CRUEL_00, +BEAUTIFUL_00, +CRIME_00, +TRIAL_00, +PUNISH_00, etc.), are used in FunGramKB as defining units which enable the construction of meaning postulates for basic concepts and terminals, as well as taking part as selectional preferences in thematic frames.
c. *Terminals* (e.g., $ABOUND_00, $EMBEZZLE_00, $VICTIM_00, $FELONY_00, $DISHONEST_N_00, etc.) are headed by the symbol $. The borderline between basic concepts and terminals is based on their definitory potential to take part in meaning postulates. Hierarchical structuring of the terminal level is practically non-existent.
d. *Subconcepts*, preceded by -, offer slight modifications as to the semantic nature of conceptual arguments, e.g., –SAVE.

Basic and terminal concepts in FunGramKB are semantically defined on the basis of conceptual schemata: thematic frames and meaning postulates. Each conceptual unit is typically analyzed in terms of one thematic frame and one meaning postulate. Thematic frames (TF) represent the number and type of the semantic participants and their corresponding conceptual selectional preferences (Jiménez-Briones & Luzondo-Oyón, 2011), whereas meaning postulates (MP) comprise a number of connected predications or events (e1, e2…) that specify compulsory and optional conceptual arguments. Meaning postulates employ COREL (Conceptual Representation Language), a machine-tractable metalanguage for the representation of lexico-conceptual knowledge (Periñán-Pascual & Mairal-Usón, 2010).

Figure 2 and Table 1 depict how the concept +OBTAIN_00 is represented in the ontological hierarchy, along with its TF and MP. An approximate translation into natural language is given afterwards:



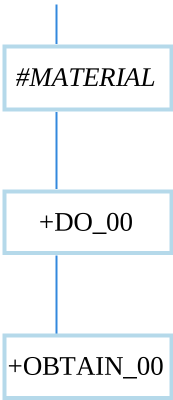



**Figure 2.** +OBTAIN_00 in the ontological hierarchy

**Table 1.** Conceptual grid of +OBTAIN_00 in FunGramKB

| | |
|---|---|
| **TF:** | (x1: +HUMAN_00)Theme<br>(x2: +ARTEFACT_00)Referent |
| **MP:** | +(e1: +DO_00 (x1)Theme (x3)Referent (f1: (e2: +HAVE_00 (x1)Theme (x2)Referent))Result)<br>+(e3: past n +HAVE_00 (x1)Theme (x2)Referent) |
| **English lexical units:** | *acquire, get, obtain* |

The ontological hierarchy provides relations of inheritance, from the most general to the most concrete concept. We can observe that +OBTAIN_00 is defined in terms of +DO_00, its genus or superordinate basic concept. The conceptual grid provides more fine-grained conceptual-cognitive information in the TF and MP. The TF specifies that the entity that obtains something is a person, and the obtained thing is an artefact, i.e., a human-made object. In the MP, +OBTAIN_00 is defined along these lines:

e1: A human being does something.
e2: As a result, the human being gets an artefact.
e3: In the past, the human being did not have it.

The English lexical units that represent these conceptual specifications are *acquire*, *get*, and *obtain*. Each of these lexical units inherits the conceptual properties of +OBTAIN_00. However, each can also specify its own lexical collocate behavior. In other words, the lexical collocates that populate a particular syntactic slot in a given verb (e.g. *get*) might not coincide with those in other verbs (e.g. *obtain* and *acquire*). In any case and under any circumstance, a conceptual unity underlies all lexical units, in that the theme should be +HUMAN_00 (i.e. the obtainer) and the referent +ARTEFACT_00 (i.e. the obtainee). The conceptual properties are not applicable to the English language only, but in principle to any other natural language.

The FunGramKB subontologies or satellite ontologies are still under development. One that is closely related to the conceptual scenario of THEFT is the *Globalcrimeterm* subontology (Felices-Lago, 2015). It contains specialized concepts belonging to the expert area of criminal law[4]. Some theft-related concepts were already added to the core ontology, such as $EMBEZZLE_00 and a mirror concept for +STEAL_00, resulting from the analysis of the *Globalcrimeterm* corpus. This corpus was compiled with legal texts from highly regarded sources in the criminal law field (Felices-Lago, 2016).

The ontology building and refinement processes of the FunGramKB Core Ontology relies on lexicographic resources such as dictionaries, thesauri, corpora, and other lexical databases for the extraction of rich semantic knowledge on the basis of meaning definitions (Faber & Mairal-Usón, 2018) and/or corpus-based collocational information (Jiménez-Briones & Luzondo-Oyón, 2011), ideally from as many different languages as possible. Our contribution departs from both, meaning definitions and collocational information of stealing verbs, in order to integrate our corpus-based findings into the Ontology through a new corpus-based method based on automatic collocate extraction. This process facilitates the extraction of semantic knowledge, probably in a more consistent way.

## 3. Methodology

Along the following lines, we present our proposed corpus-based methodology for an automatic extraction of lexical collocates. We depart from Fernández-Martínez & Faber (2020) comprehensive analysis of verbs of stealing. These were originally taken from the

[4] Some previous legal ontologies based on crimes (in general) or crimes connected with fraud paved the way for the design of *Globalcrimeterm*, as it was the case of Asaro et al. (2003), Leary et al. (2005) and some references in Valente (2005) and Sartor et al. (2011).

Lexical Grammar Model (LGM) hierarchies (Faber & Mairal-Usón, 1999): *steal*, *rob*, *hold up*, *stick up*, *thieve*, *embezzle*, *purloin*, *pilfer*, *pinch*, *filch*, *lift*, *nick*, *swipe*, *rustle*, *clean out*, *make off with*, *plunder*, *pillage*, *despoil*, *loot*. The syntactic constituents relevant for this study are those that represent the semantic valency of stealing verbs: subjects, direct objects, and oblique objects for the semantic roles perpetrator, goods, victim, and source.

### **3.1.** Presenting the Lexical Collocate Extractor (LeCoExt) tool

Our automatic corpus-based study of selection preferences is based on the data offered by *Sketch Engine*, which is a large database of corpora that incorporates user-friendly tools for accessing and retrieving information from their corpora as well as creating new corpora (Kilgarriff et al., 2014). For instance, for words and phrases *Sketch Engine* offers functionalities for the analysis of examples in actual discourse, collocational patterns, semantic relations of synonymy, or translation lookups. *Sketch Engine* can also generate lists of the most frequent tokens in a corpus, extraction of key words and terms and their translation in other languages, calculate n-grams, recognize neologisms, tag texts with part-of-speech tags, etc. One of the *Sketch Engine*'s main functionalities is the Word Sketch tool (Kilgarriff et al., 2010; McCarthy et al., 2015). A word sketch is defined as "an automatic corpus-derived summary of a word's grammatical and collocational behavior" (Kilgarriff et al., 2010, pp. 372). In the case of verbal predicates, word sketches can provide information about the most frequent lexical collocates that perform specific syntactic functions in the clause such as subject and object. These syntactic constituents receive the name of grammatical relations or 'gramrels'.

Our LeCoExt command-line tool was coded in the programming language C# from the .NET framework. It works by requesting HTTP access to the *Sketch Engine* API.[5] After the authentication step, we specify the parameters for automatic Word Sketch searches using object-oriented Corpus Query Language (CQL) commands (Jakubíček et

[5] Further documentation details can be found on https://www.sketchengine.eu/documentation/api-documentation/ [last accessed 16 May 2019]

al., 2010). For instance, we can specify the corpora used, the lemmas, their part of speech, the number of items appearing for each gramrel, and a minimum frequency score. We select the lemmas corresponding to stealing verbs and the verb part-of-speech. Then, we specify, since verbs are the target, the specific or relevant gramrels with reference to the stealing verb in question: SUBJECT OF X, OBJECT OF X, X FROM, and/or X OF. The minimum frequency score is set to 0.0. The corpora selected are the BNC Corpus (2007), the English Web 2013 and 2015 corpora (enTenTen13 and enTenTen15, see Jakubíček et al., 2013), and the Timestamped JSI 2014-2019 English Web Corpus (Bušta & Herman, 2017)[6]. Figure 3 summarizes the sequential pipeline performed by our tool through a flowchart:

[6] Monthly updates aggregate new labeled data to the Timestamped JSI corpus, which means that the collocates and their frequency scores presented on this article, when replicated, might be slightly different. Further information about the in-progress status of the corpus can be found on https://www.sketchengine.eu/jozef-stefan-institute-newsfeed-corpus/ [last accessed 12 November, 2019]

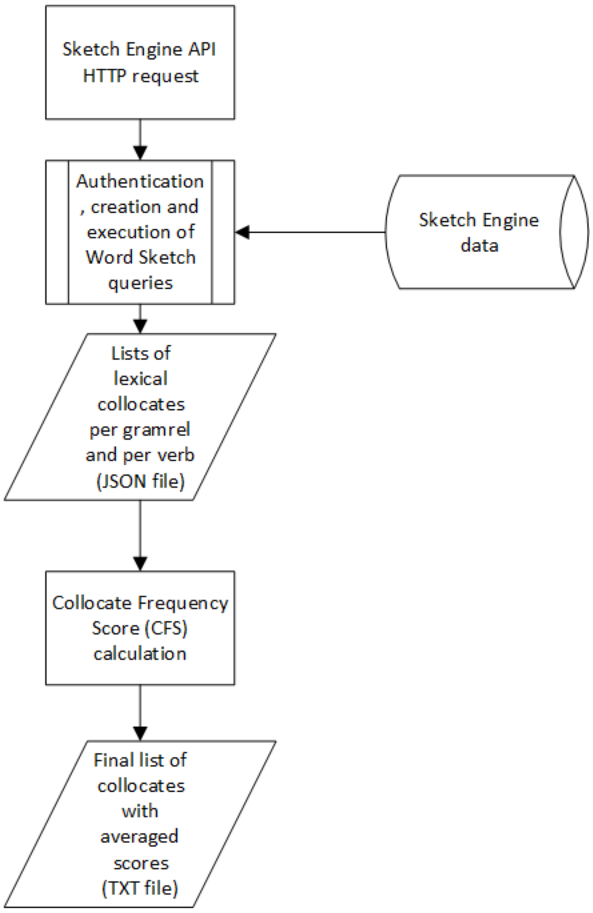



**Figure 3.** A flowchart of the LeCoExt tool pipeline

Following the initial configuration step, our tool automatically retrieves the gramrels and provides lists of a maximum allowed length of 25 lexical collocates with their corresponding logDice frequency score (Rychlý, 2008). To avoid adding erroneous collocates from corpus noise such as random combination of letters and numbers, we apply a regular expression (regex) rule that discards them. After the retrieval of collocates, the tool can also process some calculations automatically to obtain an average of each lexical collocate. First, it calculates the means of each lexical collocate according to a) the number of corpora in which they appear, and b) the number of verbs, besides the verb in question, where the lexical collocate appears. This results in the average frequency score formula that can be seen in figure 4, where epsilon is the sum of each logDice salience score $j$ for one lexical collocate, $n$ is the number of corpora used for the search, and $k$ is an integer variable whose default value is 0 that represents the number of verbs, besides the current one, from which the lexical collocate is obtained.

$$Average\ frequency\ score = \frac{\sum_{i=0}^{j}}{n + k}$$

**Figure 4.** Average frequency score

The *k* variable is introduced because it provides balance in numeric data for the calculation of the mean. If we have a collocate X that appears in the set number of *n* corpora with one verb, but it also appears with other verbs, we need the *k* variable to get a correct mean estimate.

Another important variable to take into account is the frequency of lexical collocates per verbs. In other words, we measure the number of verbs in which the lexical collocate appears divided by the total number of verbs that are used for the search. This formula is a weighting factor (see figure 5), which is multiplied to the previous calculation to penalize those collocates that, despite having a large frequency score with one particular stealing verb, are missing or have a very low frequency score with other verbs. Its maximum value is 1:

$$Weighting\ factor = \frac{number\ of\ \ verbs\ for\ lexical\ collocate}{total\ number\ of\ \ verbs}$$

**Figure 5.** Weighting factor

Thus, our Collocate Frequency Score (CFS), based on previous calculations, is in Figure 6:

$$Collocate\ frequency\ score = Average\ frequency\ score \cdot \ weighting\ factor$$

**Figure 6.** Collocate Frequency Score

An advantage of this formula over the average frequency score is that we can get a more accurate estimate of the collocational behavior of a group of verbs that make up one lexical domain in terms of the frequency of lexical collocates, not only per corpus and individual verbs, but also per the whole group of verbs. In this way, we also avoid wrong or erroneous lexical collocates —because of a malfunctioning part-of-speech tagger or syntactic chunker— to appear high on the lists, since their presence in one verb is merely anecdotal. It is also useful to minimize the presence of collocates that appear in verbs corresponding to verbal senses other than those related to THEFT.

For a practical application of the proposed formulas, let us consider the collocate *bandit*, one that can be a prototypical perpetrator in the subject argument slot of stealing verbs. Under the aforementioned set parameters for the automatic search, and considering all stealing verbs from the THEFT domain, we obtain that, in the 4 corpora used, *bandit* has an approximate logDice salience score of 4.88, 5.66, and 5.30 for the lemma *loot;* 4.1 for the lemma *pilfer*; 4.05, 4.73, and 3.43 for *pillage*; 4.71, 4.74 for *plunder*; 6.92, 7.51, and 7.78 for *rob*; 5.61 for *steal*; and 5.5 for *swipe*. The exact sum of all scores is 74.86. This number is divided by the sum of 4, the number of corpora, plus *k*, whose value is 6, i.e., the number of total verbs in which bandit appears as subject minus the first one. That gives 7.486, which corresponds to the balanced mean of all logDice salience scores. The weighting factor is now applied to the previous calculation to provide an even more balanced numeric estimate of the weight of the collocate *bandit* in stealing verbs. The weighting factor calculation is 7, the number of verbs in which the collocate appears, divided by 21, the total number of stealing verbs used in the search. The resulting value, 0.3333, is multiplied to the previously calculated average frequency score, 7.486, giving us approximately a CFS of 2.4953.

In addition to this global search, we perform local searches for specific lexical units of the THEFT domain that, according to their meaning definitions (Fernández-Martínez & Faber, 2020), specify differentiating parameters that make them different enough from their superordinate term *steal*, and which may potentially serve to create or refine subconcepts in the FunGramKB ontology. The verbs in question are *embezzle*, *shoplift*, *rustle*, and plundering verbs (e.g., *plunder*, *pillage*, *loot*, and *despoil*) as a semantically-cohesive subgroup of stealing verbs.

### **3.2.** FunGramKB conceptual categorization and specifications

The next step involves mapping the resulting lexical collocates of each search onto FunGramKB conceptual categories before refining or creating new concepts for the THEFT scenario. This was done by manually checking the *FunGramKB* concepts in the core ontology for each group of semantically-related lexical collocates. For instance, to save time, we do not perform individual searches of collocates that are related to the *money* (e.g., *cash*, *dollar*, *coins*, etc.) collocate since they are semantically interrelated. Instead, we only perform one search using the most representative or prototypical word which, in this case, is *money*. Table 2 provides some FunGramKB conceptual categories together with some examples of lexical collocates that represent the lexical realizations of these categories:

**Table 2.** Some relevant FunGramKB conceptual categories for the THEFT conceptual scenario

| **FunGramKB conceptual categories** | **Lexical collocates** |
|---|---|
| +ANIMAL_00 | *cat, dog, horse, fox, cattle, livestock* |
| +ARTEFACT_00 or +ARTIFICIAL_OBJECT_00 | *artworks, painting, sculpture, money, cash, riches, wealth, pounds, dollars, euros, land, estate, property, booty. treasure, wallet, purse, bag, wristband, lipstick, nail polish, razor, money, perfume, book, passport, ticket, card, hammer, drill, drumstick, car, bike, auto, caravan, gun, firearm, rifle, TV, phone, radio, laptop, stereo, clothes, t-shirt, underwear, skirt, trousers, handkerchief, jewelry, necklace* |
| +CATTLE_00 | *cattle, livestock, cow* |
| +NATURAL_OBJECT_00 and +SUBSTANCE_00 | *meat, fruit, vegetables, grains, legumes, milk, water, gas, petrol* |
| +PLACE_00 | *shop, supermarket, church, bank, garden, countryside, farm, house* |
| $THIEF_00 | *thief, criminal, gang, band, government, state, multinational* |
| $VICTIM_00 | *human, animal, tourist* |

For the implementation of our findings into FunGramKB, we use the in-built representational language COREL to update or create concepts in the ontology. At this stage, we need to know which theft-related concepts are already embedded in the ontology. For that purpose, we search with some stealing-verb keywords in the ontology to obtain the extant concepts and their conceptual patterns and definitions. From this search, we gather that there is a general concept +STEAL_00 that groups all stealing verbs. We also find the terminal concept $EMBEZZLE_00, yet we cannot observe any reference to any other stealing concepts or subconcepts. All the conceptual specifications of new theft-related concepts are sketched out in the following section after the automatic search of collocates.

## 4. Results and discussion

### 4.1. Extraction of semantic knowledge

The command-line window provides output lists of lexical collocates for the semantic frame elements of perpetrator, goods, and source respectively for the THEFT domain as a whole[7]. In tables 3, 4, and 5, we can observe the 25 most frequent lexical collocates for each semantic argument with their CFS:

**Table 3.** List of collocates from the perpetrator frame (subject gramrels)

| Collocate | CFS |
|---|---|
| thief | 5,42274508055519 |
| robber | 2,92366666793823 |
| bandit | 2,49533336162567 |
| nazis | 2,24920638402303 |
| burglar | 2,14539682297480 |
| pirate | 2,01333334332421 |
| gang | 1,72351190376849 |
| crook | 1,64380952290126 |
| criminal | 1,39015872516329 |
| looter | 1,33166666257949 |
| cattle | 1,10690479051499 |
| hacker | 1,08571430777206 |
| invader | 0,995476200467064 |
| raider | 0,968163263230097 |
| breeze | 0,799809528532482 |
| suspect | 0,764952378045945 |
| finger | 0,761666669732048 |
| viking | 0,695952381406512 |
| brow | 0,620190484183175 |
| marauder | 0,594761905216035 |
| knife | 0,578333343778338 |
| crusaders | 0,551666657129923 |
| nazi | 0,550095240275065 |
| thug | 0,550095231192453 |
| man | 0,545850345066615 |

[7] The search was performed on May 23, 2019.

In Table 3, there are a few lexical collocates such as *cattle*, *finger* and *brow* whose presence can be explained as emerging from a malfunction in the part-of-speech tagging or syntactic chunking, and/or other meanings of stealing verbs. For instance, *cattle stealing* is a frequent collocational pattern that appears in the corpora as subjects of some stealing verbs. *Brow* and *finger* probably come from the multi-word combinations *brow lift* and *lift finger*. If it were not by the weighting factor, these collocates would have appeared higher on the list. We also want to emphasize a strong axiological component in the semantic type of perpetrator agents (Felices-Lago, 2014). Many of these collocates are loaded with negative connotations, since a stealing act is usually an illegal activity that results in crime. This means that we do not expect positively-loaded words such as *nun* or *baby* to be perpetrators of criminal acts. Also, despite not being represented high on the list, we find examples of animals as perpetrator agents of stealing acts. All this semantic knowledge can thus be linked to the FunGramKB conceptual category $THIEF_00, a specific type of criminal.

**Table 4.** List of collocates from the goods frame (object gramrels and of-object gramrels)

| Collocate | CFS |
|---|---|
| treasure | 2,66770564322864 |
| wallet | 2,60966665744781 |
| purse | 2,0563333183527 |
| treasury | 1,99346937452044 |
| coffer | 1,941179485565 |
| valuable | 1,8086309447175 |
| finger | 1,60880951654343 |
| countryside | 1,40850663239183 |
| cash | 1,19904758627453 |
| tomb | 1,08380951767876 |
| monastery | 1,05842565061361 |
| antiquity | 0,999988657160797 |
| card | 0,966428575061616 |
| phone | 0,93214285941351 |
| money | 0,887142851239159 |
| possession | 0,820952376540826 |
| fund | 0,818775512734238 |
| belonging | 0,818095241274152 |
| spree | 0,770476182301839 |
| good | 0,714583329945093 |
| bike | 0,66533332098098 |

| | |
|---|---|
| base | 0,665142858596075 |
| pirate | 0,640476187070211 |
| chin | 0,623428571791876 |
| riches | 0,616921776276327 |

Here, the most typical goods are valuable goods. One of the most characteristic semantic category is that of assets, represented in the ontology as +MONEY_00. Also, we find many other categories here such as vehicle, electronic device, container, etc. All these represent the +ARTEFACT_00 category. Though not exemplified on this shortened list, animals such as *cat* or *cattle* can also be stolen, so +ANIMAL_00 is another likely category. Other categories for goods are +NATURAL_OBJECT_00 (e.g., *meat*, *fruit*, *minerals*, etc.), +SUBSTANCE_00 (e.g., *drugs*) and even +HUMAN_00 (e.g., *a baby*). Some outliers also figure on the list, such as *tomb* or *monastery*. This can be explained on the grounds that some verbs (i.e., plundering verbs) exhibit complementary syntactic patterns in which their objects refer to the locational origin of the plundering act.

**Table 5**. List of collocates from the source frame (from-object and object gramrels)

| Collocate | CFS |
|---|---|
| treasury | 3,10078230358305 |
| store | 2,79957672095353 |
| shop | 2,34085318834063 |
| coffer | 1,941179485565 |
| tomb | 1,70725622060212 |
| countryside | 1,40850663239183 |
| treasure | 1,29551021744605 |
| monastery | 1,05842565061361 |
| employer | 1,01833333855584 |
| antiquity | 0,999988657160797 |
| room | 0,831020413612833 |
| victim | 0,82171428771246 |
| bank | 0,782095223381406 |
| car | 0,769761914298648 |
| purse | 0,716428575061616 |
| fund | 0,713356416907788 |
| account | 0,705238103866577 |
| residence | 0,664742060529402 |
| riches | 0,616921776276327 |
| museum | 0,609142848423549 |
| artefact | 0,600364430112131 |
| temple | 0,585855508970527 |

| | |
|---|---|
| pharmacy | 0,583568595434262 |
| valuable | 0,581714284987677 |
| page | 0,572761908031645 |

Some outliers that mostly come from plundering verbs figure here. Anyway, we can still observe that the source is +PLACE_00 or +HUMAN_00. To be more specific, instead of +HUMAN_00, we select the terminal concept $VICTIM_00, which refers to the target of a criminal activity. The places where stealing takes place are actually metonymic projections of the type PLACE FOR PEOPLE, since the owner cannot be a location, but a victim. This also applies to *purse*, which is a container, not a place per se, but a metonymic extension of the type PLACE FOR CONTAINER. Thus, the locational semantics of stealing verbs can be claimed to specify where the stealing act occurs (e.g., *a bank*, *a museum*, etc.) or where the goods are contained (e.g., *in a purse*, *in a wallet*, etc.). The locational semantic aspect must be mentioned in the MP of the THEFT scenario through an optional argument in the event predicate that takes +PLACE_00 and/or +CONTAINER_00 for a location argument.

The following tables represent the collocates CFS of the verbs *shoplift*, *embezzle*, *rustle*, and plundering verbs. This time we provide the ten most frequent collocates for each frame. For reasons of space, we offer the approximate means of each collocate:

**Table 6**. List of collocates from *shoplift*

| **Perpetrator frame** | | **Goods frame** | | **Source frame** | |
|---|---|---|---|---|---|
| **Collocates** | **CFS** | **Collocates** | **CFS** | **Collocates** | **CFS** |
| misdemeanor | 6,60 | spree | 5,19 | walmart | 5,93 |
| felony | 5,96 | trespassing | 4,64 | apparel | 4,40 |
| re-offence | 2,89 | larceny | 4,56 | store | 3,75 |
| cctv | 2,43 | vandalism | 4,47 | wal-mart | 3,69 |
| walmart | 2,19 | burglary | 4,35 | supermarket | 3,27 |
| third-degree | 2,13 | theft | 4,33 | target | 3,13 |
| fourth-degree | 1,97 | incident | 3,96 | macy | 2,32 |
| second-offense | 1,78 | offence | 3,74 | sephora | 2,10 |
| inocencio | 1,76 | suspect | 3,70 | boots | 2,05 |

| | | | | | |
|---|---|---|---|---|---|
| third-offense | 1,73 | arrest | 3,27 | penney | 1,99 |

We can observe how a malfunctioning syntactic part-of-speech tagger and chunker hinders any objective evaluation of the semantics of *shoplift*. Since none of the ten most frequent collocates are actual collocates of the verb but the gerund as a noun, we need to have a look at the full list obtained from *shoplift*. There, we find collocates such as *cologne*, *ring*, *DVDs*, *seafood*, *vitamin*, *cosmetic*, and *cigarillo* for the goods frame. In other words, shoplifting involves stealing +ARTEFACT_00, +SUBSTANCE_00, and +NATURAL_OBJECT_00. The perpetrator is always +THIEF_00 and the underlying semantics of shoplift specify the location +SHOP_00, but not the victim.

**Table 7**. List of collocates from *embezzle*

| Perpetrator frame | | Goods frame | | Source frame | |
|---|---|---|---|---|---|
| **Collocates** | **CFS** | **Collocates** | **CFS** | **Collocates** | **CFS** |
| bookkeeper | 5,60 | fund | 5,27 | employer | 4,35 |
| accountant | 4,29 | ruble | 4,44 | fund | 2,86 |
| treasurer | 3,68 | rouble | 3,98 | treasury | 2,46 |
| out-worker | 2,54 | yuan | 3,87 | midwinter | 2,38 |
| ablyazov | 2,05 | billion | 3,58 | bank | 2,34 |
| sandinista | 1,98 | million | 3,57 | company | 2,32 |
| picker | 1,84 | sum | 3,45 | firm | 2,31 |
| amount | 1,83 | money | 3,36 | berhad | 2,11 |
| saldivar | 1,79 | monies | 3,16 | organization | 2,05 |
| clique | 1,75 | fcfa | 2,99 | petrobras | 1,91 |

The prototypical cognitive-conceptual pattern of embezzling is that of a worker of a company or organization (i.e., +WORKER_00) that takes care of monetary assets that belong to the company —or more precisely, the owners or founders of the company (i.e., +EMPLOYER_00)—. Thus, the conceptual category for goods is +MONEY_00. Though unrepresented on the lists, we also find very infrequent examples in the manual concordances of *embezzle* which are further examined in actual discourse to find out whether they are corpus noise. These goods other than *money* are *ibuprofen* (i.e., +SUBSTANCE_00), *reptile* (i.e., +ANIMAL_00), a *bible* (i.e., ARTEFACT_00) and *sapphire* (i.e., +NATURAL_OBJECT_00), among others. This can be justified if embezzling is understood as illegally taking any kind of goods as long as these are owned by an institution. Since our interest is not only to capture the most frequent conceptual categories but to cover the full scope of conceptual categories, it makes sense to include those in the FunGramKB conceptual grid.

**Table 8**. List of collocates from *rustle*

| **Perpetrator frame** | | **Goods frame** | | **Source frame** | |
|---|---|---|---|---|---|
| **Collocates** | **CFS** | **Collocates** | **CFS** | **Collocates** | **CFS** |
| breeze | 7,65 | cattle | 5,66 | breeze | 1,76 |
| cattle | 7,07 | leave | 5,12 | bush | 1,15 |
| faint | 6,81 | frond | 5,10 | right | 0,64 |
| wind | 5,50 | grass | 5,06 | farm | 0,53 |
| leave | 5,09 | Jimmy | 4,97 | passage | 0,48 |
| livestock | 4,13 | silk | 4,71 | outside | 0,42 |
| sheep | 3,90 | noise | 4,44 | forest | 0,31 |
| skirt | 3,60 | sound | 4,28 | | |
| leaf | 3,20 | feather | 4,17 | | |
| grass | 3,16 | foliage | 3,92 | | |

In this case we are attesting how other meanings flood the list with collocates different from the ones related to the stealing sense. On top of that, a malfunctioning processing of the concordances has resulted in a grammatically-incorrect characterization of the collocational behavior of *rustle*. Looking at the whole list, we can see that rustling is a stealing act performed by a +THIEF_00 that takes farm animals (i.e., +CATTLE_00) from a +FARMER_00 in a farm or barn (i.e., +FARM_00).

**Table 9**. List of collocates from plundering verbs: *plunder, pillage, despoil, loot*

| **Perpetrator frame** | | **Goods frame** | | **Source frame** | |
|---|---|---|---|---|---|
| **Collocates** | **CFS** | **Collocates** | **CFS** | **Collocates** | **CFS** |

| | | | | | |
|---|---|---|---|---|---|
| looter | 6,99 | treasury | 8,39 | treasury | 8,39 |
| nazis | 6,93 | countryside | 7,97 | countryside | 7,97 |
| invader | 5,23 | tomb | 7,42 | tomb | 7,42 |
| robber | 4,86 | treasure | 7,06 | Treasure | 7,06 |
| pirate | 4,75 | coffer | 6,97 | Coffer | 6,97 |
| bandit | 4,68 | monastery | 6,13 | monastery | 6,13 |
| marauder | 3,12 | antiquity | 5,80 | antiquity | 5,80 |
| crusaders | 2,90 | riches | 3,62 | Riches | 3,62 |
| nazi | 2,89 | artefact | 3,49 | Artefact | 3,49 |
| vandal | 2,77 | artifact | 3,19 | Artifact | 3,19 |

Plundering verbs form a neat subgroup of stealing verb that share the same encyclopedic knowledge, i.e., stealing valuables during times of turmoil, in violent events, usually causing damage to the plundered areas. The fact that the goods frame and source frame share the same collocates and CFS is due to the complementary patterns of these verbs which admit that either the goods or the source can function as direct object or as oblique object. Anyway, it is clear that a +THIEF_00 is involved in the act of stealing, usually organized in large groups. Looking further down in the full list, we can observe that the stolen goods can be +ARTEFACT_00 (e.g., *monetary assets*, *jewels*), +SUBSTANCE_00 (e.g., *petrol*), +NATURAL_OBJECT_00 (e.g., *diamonds*, *rubies*), and even +CATTLE_00 (e.g., *livestock*). The source is $VICTIM _00, and the place where plundering happens (+PLACE_00), as well as the aforementioned encyclopedic knowledge, must be implemented in the conceptual grid of a new terminal concept $PLUNDER_00.

### **4.2.** Implementation and refinement of findings into the FunGramKB Core Ontology

Following the COHERENT methodology (Periñán-Pascual & Mairal-Usón, 2011) and the guidelines for the creation of concepts (Jiménez-Briones & Luzondo-Oyón, 2011), we explore the refinement and implementation of the conceptual meaning of stealing verbs into the FunGramKB Core Ontology.

The semantic knowledge extracted from our corpus findings and the Lexical Grammar Model hierarchy of stealing verbs helps us configure the conceptual specifications in terms of TFs and MPs. The FunGramKB Core Ontology already encapsulates two concepts belonging to the conceptual scenario of theft, namely, the basic concept +STEAL_00 and the terminal concept $EMBEZZLE_00, which saves us the conceptualization and hierarchization phases of the COHERENT methodology. However, we detect various problems and discrepancies related to the definition and location of these concepts in the ontology. We thus proceed to the third and fourth phases of the COHERENT methodology, particularly, those of remodeling and refinement.

First of all, we notice that +STEAL_00 and $EMBEZZLE_00 are differently placed in the ontology: the genus of $EMBEZZLE_00 is +GIVE_00 and the genus of +STEAL_00 is +OBTAIN_00 (Felices-Lago, 2016, pp. 132-133). However, if *embezzle*, as a lexical unit, is subordinate to *steal*, the terminal concept $EMBEZZLE_00 should be defined on the basis of the basic concept +STEAL_00 and not any other. Then comes the refinement of the TF and MP of +STEAL_00 on the basis of our corpus findings and the semantic knowledge extracted.

The lexical collocates and their corresponding conceptual categories show that *steal* and its subordinate units offer a much richer semantic network than that displayed in the current version of the Ontology. In light of our findings, Table 10 is the proposed refinement of the concept +STEAL_00, where changes are marked in red:

**Table 10.** Conceptual grid of +STEAL_00

| | |
|---|---|
| **TF:** | (x1: +THIEF_00)Theme<br>(x2: +ARTEFACT_00 ^ +NATURAL_OBJECT ^ +SUBSTANCE_00 ^ +HUMAN_00 ^ +ANIMAL_00)Referent |
| **MP:** | +(e1: +OBTAIN_00 (x1)Theme (x2)Referent (f1: +VICTIM_00)Origin (f2: +PLACE_00 ^ +CONTAINER_00)Location (f3: +CRIME_00)Result) |
| **English lexical units:** | *steal, rob, hold up, stick up, thieve, embezzle, purloin, pilfer, pinch, filch, lift, nick, swipe, rustle, clean out, make off with, plunder, pillage, despoil, loot* |

As for +EMBEZZLE_00, in light of the present observations, we present a refined conceptual grid (Table 11):

**Table 11.** Conceptual grid of $EMBEZZLE_00

| | |
|---|---|
| **TF:** | (x1: +WORKER_00)Theme (x2: +ARTEFACT_00 ^ +NATURAL_OBJECT_00 ^ +SUBSTANCE_00 ^ +ANIMAL_00)Referent |
| **MP:** | +(e1: +STEAL_00 (x1)Theme (x2)Referent (f1: +EMPLOYER_00)Origin (f2: +HUMAN_00)Goal (f3: (e2: +USE_00 (f2)Theme (x2)Referent)Purpose)) (f4: +PLACE_00)Location |
| **English lexical units:** | *embezzle* |

Also, we can sketch some lexico-semantic differentiating parameters from the selection preferences and the underlying semantics of verbs such as *shoplift*, *rustle*, and plundering verbs that make them more specific than +STEAL_00 in terms of their conceptual selection preferences and their encyclopedic knowledge. This means that these new terminal concepts inherit the main properties of the basic concept +STEAL_00 but present differentiae in their TFs and MPs.

The proposed concepts are $SHOPLIFT_00, $RUSTLE_00, and $PLUNDER_00. A sketch of their conceptual specifications is offered in tables 12, 13, and 14, respectively:

**Table 12.** Conceptual grid of $SHOPLIFT_00

| | |
|---|---|
| **TF:** | (x1: +THIEF_00)Theme<br>(x2: +ARTEFACT_00 ^ +SUBSTANCE_00 ^ +NATURAL_OBJECT_00)Referent |
| **MP:** | +(e1: +STEAL_00 (x1)Theme (x2)Referent (f1: +SHOP_00)Location |
| **English lexical units:** | *shoplift* |

**Table 13.** Conceptual grid of $RUSTLE_00

| | |
|---|---|
| **TF:** | (x1: +THIEF_00)Theme<br>(x2: +CATTLE_00)Referent |
| **MP:** | +(e1: +STEAL_00 (x1)Theme (x2)Referent (f1: +FARMER_00)Origin (f2: +FARM_00)Location |
| **English lexical units:** | *rustle* |

**Table 14.** Conceptual grid of $PLUNDER_00

| | |
|---|---|
| **TF:** | (x1: +THIEF_00)Theme |

| | |
|---|---|
| | (x2: +ARTEFACT_00 ^ +NATURAL_OBJECT_00 ^ +SUBSTANCE_00 ^ +ANIMAL_00)Referent |
| **MP:** | +(e1: +STEAL_00 (x1)Theme (x2)Referent (f1: $VICTIM_00)Origin (f2: +PLACE_00)Location (f3: +DAMAGE_01)Result (f3: (e2: +BE_01 (x3: +OCCURRENCE_00 ^ +PERIOD_00)Theme (x4: +VIOLENT_00)Attribute))Scene) |
| **English lexical units:** | *plunder, pillage, despoil, loot* |

The FunGramKB English lexicon can also benefit from our corpus-based study of the selection preferences of stealing verbs, incorporating the specific collocational behavior of each verb into its database. Each conceptual entry is associated to a number of lexical units across different languages. Lexical entries could thus be devised for each of the lexical units associated to each concept. The specific collocational behavior of particular verbs can be implemented in the tab LEXICAL TEMPLATE (collocations).

We must emphasize that despite the linguistically-driven character of the ontology building and refinement processes, the resulting concepts are language-independent. In other words, the concepts are universal and can be differently represented across languages. For instance, $SHOPLIFT_00 does not have a direct lexical counterpart in Spanish, but can be paraphrased as *hurtar en tienda* 'steal from shops'. Due to the universal character of concepts, we assume that all these concepts can be lexicalized or paraphrased in any human language and that all corresponding lexicalizations will be governed by the conceptual specifications and categories defined in the conceptual grids.

Having said that, we offer a few examples of how the implementation of lexical selection preferences would go about for the English lexicon:

**Table 15.** Lexical template (collocations) of *steal*

| | |
|---|---|
| **X-theme** | *thief, burglar, hacker, gang, robber, criminal, suspect* |
| **Y- theme** | *car, bike, wallet, purse, vehicle, money* |

**Table 16.** Lexical template (collocations) of *rob*

| | |
|---|---|
| **X-theme** | *bandit, thief, gunman, gang, suspect, thug, robber* |
| **Y- theme** | *bank, store, victim, pharmacy, dealer, passenger, shop* |
| **Z-referent** | *valuable, wallet, cellphone, belonging, purse, possession* |

**Table 17.** Lexical template (collocations) of *embezzle*

| | |
|---|---|
| **X-agent** | *bookkeeper, accountant, treasurer, out-worker* |
| **Y- theme** | *fund, ruble, yuan, money, rupee, euro, dollar* |

**Table 18.** Lexical template (collocations) of *shoplift*

| | |
|---|---|
| **X-theme** | *thief, teenager* |
| **Y- theme** | *merchandise, underwear, possession, cologne, ring, dvds, seafood, vitamin, cosmetic* |

**Table 19.** Lexical template (collocations) of *plunder*

| | |
|---|---|
| **X-theme** | *looter, pirate, Nazis, invader, imperialist, raider, conquistador, privateer* |
| **Y- theme** | *booty, riches, treasure, wealth, valuable* |
| **Z-referent** | *treasury, countryside, coffer, tomb, monastery* |

### **4.3.** Limitations in this contribution

In our automatic search of lexical collocates there are a number of issues that need further discussion. One of them is related to the presence of collocates in the lists that need manual evaluation and intervention for their removal. For instance, we have detected wrong collocates that appeared in the perpetrator list of stealing verbs such as *cattle, brow,* or *finger*. Moreover, we have also observed some other outliers in other lists, especially in the lexical collocate extraction of individual verbs, as in the perpetrator list of *shoplift*. At times, the same lemma appears twice on the lists, such as the word *nazi* and *nazis* for the perpetrator frame of plundering verbs. This can be explained mainly in terms of the way in which the in-built syntactic POS taggers work, since a wrongly attributed POS tag hinders the calculations and display of lexical collocates, affecting any derived application of word sketches.

Another important disrupting factor in the display of lexical collocates is linked to the lexicosemantic phenomenon of polysemy. The *Word Sketch* tool cannot differentiate lexical collocates in terms of the different verbal senses, which might affect the presentation of collocates on the lists, since we are interested in the literal senses that represent THEFT scenarios only. We have observed this in verbs with many polysemous senses such as *lift*, *pinch*, *nick*, or *swipe.* The stealing meaning of these polysemous verbs is sometimes secondary, and extracting their stealing-related collocates becomes a difficult task, unless done manually. Also, some generic stealing verbs such as *steal* and *rob* exhibit many collocational patterns in secondary senses (e.g. *steal a limelight*, *steal a show*, *steal a glance*, *rob Peter to pay Paul*, etc.). This can be problematic because metaphoric uses and senses are ubiquitous. That is why we devise the weighting factor, as it dramatically reduces the chances of these collocates showing up on the lists.

We believe that an improved POS tagger and syntactic chunker together with an in-built word-sense disambiguation system could help to organize collocates, not only according to gramrels, but also according to their meanings. This would tremendously help to achieve better accuracy in collocate extraction. Since for the time being the presence of wrong collocates in the lists cannot be avoided, any collocate extraction process demands proceeding with caution when interpreting the mined collocate lists. Some manual intervention thus becomes necessary in order to discard outliers and intruders.

## 5. Conclusion

To the best of our knowledge, LeCoExt, the proposed working prototype of a collocate extraction tool, is the first attempt in developing an automatic method for a comprehensive corpus-based study of the selection preferences of individual or groups of verbs. The direct application of this tool in the corpus-based building and refinement processes of FunGramKB shows its performance efficiency, which could facilitate knowledge engineers, terminologists, and linguists alike the task of extracting rich semantic knowledge from a handful of corpora. The presented tool saves us a tremendous amount of time in the labor-intensive task of traditional corpus-based methods and searches. In only a few seconds and at a click away, we can obtain all this massive amount of lexical knowledge without manual intervention. We have offered a case study of the tool being applied to stealing verbs for the extraction of the conceptual-cognitive patterns embedded in their lexical realizations and implement this knowledge into the FunGramKB Core Ontology.

In future research, the functionalities of the tool could be expanded. For instance, it could display any other gramrels of choice and offer more personalized and customized searches for other *Word Sketch* searches and functionalities. Moreover, it could be applied to many other semantically-related verbal domains (e.g., the POSSESSION domain), and across many languages, which could be useful for translation and cross-linguistic studies. We also leave to future research the development and implementation of an automatic FunGramKB conceptual categorizer of collocates into FunGramKB conceptual categories, thus further alleviating the task of ontology population and refinement.

**Funding Information**

This article is based on the R&D Project within the framework of the "Programa Operativo FEDER Andalucía 2014-2020 (code B-HUM177-UGR18)" and funded by Junta de Andalucía, Spain.

## 6. References

Asaro, C., Biasiotti, M. A., Guidotti, P., Papini, M., Sagri, M. T., & Tiscornia, D. (2003). A domain ontology: Italian crime ontology. In *Proceedings of the ICAIL 2003 Workshop on Legal Ontologies & Web based legal information management*, 1-7.

Berman, R. (1982). On the Nature of 'Oblique' Objects in Bitransitive Constructions. *Lingua*, 56(2), 101-125.

Boas, H. (2013). Frame Semantics and Translation. In A. Rojo & I. Ibarretxte-Antunano (Eds.), *Cognitive Linguistics and Translation* (pp. 125-158). Berlin/New York: Mouton de Gruyter.

British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available at http://www.natcorp.ox.ac.uk/ [last accessed 15 May 2019]

Bušta, J., & Herman, O. (2017). JSI Newsfeed Corpus. In *The 9th International Corpus Linguistics Conference*, University of Birmingham, 25-28 July 2017.

Dux, R. (2018). Frames, Verbs, and Constructions: German Constructions with Verbs of Stealing. In A. Ziem & H. Boas (Eds.), *Approaching German Syntax from a Constructionist Perspective* (pp. 367-405). Berlin/New York: Mouton de Gruyter.

Faber, P. & Mairal-Usón, R. (1999). *Constructing a Lexicon of English Verbs*. Berlin: Mouton de Gruyter.

Faber, P. & Mairal-Usón, R. (2018). A Conceptually-Oriented Approach to Semantic Composition in RRG. In R.D. Van Valin (Ed.), *The Cambridge Handbook of Role and Reference Grammar*. Cambridge: Cambridge University Press.

Felices-Lago, Á. (2014). The emergence of axiology as a key parameter in modern linguistics. In *Evaluation in Context*, G. Thompson and L. Alba-Juex (eds), 27-46. Jon Benjamins.

Felices-Lago, Á. (2015). Foundational considerations for the development of the Globalcrimeterm subontology: A research project based on FunGramKB. *Onomazéin*, 31(1): 127-144.

Felices-Lago, Á. (2016). The Process of Constructing Ontological Meaning Based on Criminal Law Verbs. *Círculo de Lingüística Aplicada a la Comunicación*, 65: 109-148.

Fernández-Martínez, N.J. & Faber, P. (2020). Who stole what from whom? A corpus-based, cross-linguistic study of English and Spanish verbs of stealing. *Languages in Contrast*, 20(1): 107-140.

Fillmore, C. & Baker, C. (2010). A Frames Approach to Semantic Analysis. In B. Heine & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (pp. 313-340). New York: Oxford University Press.

Gangemi, A., Sagri, M. & Tiscornia, D. (2005). A Constructive Framework for Legal Ontologies. In V.R. Benjamins et al. (Eds.), *Law and the Semantic Web* (pp. 97-124). Berlin: Springer.

Goldberg, A. (2010). Verbs, Constructions and Semantic Frames. In M. Rappaport-Hovav, E. Doron and I. Sichel (Eds.), *Syntax, Lexical Semantics and Event Structure* (pp. 39-58). Oxford: Oxford University Press.

Jakubíček, M., Kilgarriff, A., McCarthy, D. & Rychlý P. (2010). Fast Syntactic Searching in Very Large Corpora for Many Languages. *PACLIC*, 741-747.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen Corpus Family. *Seventh International Corpus Linguistics Conference CL*, 125-127.

Jiménez-Briones, R. & Luzondo-Oyón, A. (2011). Building Ontological Meaning in a Lexico-conceptual Knowledge Base. *Onomázein*, 23, 11-40.

Kilgarriff, A., Vojtěch, K., Krek, S., Srdanovič, I. & Tiberius C. (2010). A Quantitative Evaluation of Word Sketches. *Proceedings of the 14th EURALEX International Congress*, 372-379.

Kilgarriff, A. & Renau, I. (2013). esTenTen, a Vast Web Corpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences*, 95, 12-19.

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, 7-36. Available at http://www.sketchengine.co.uk [last accessed 28 December 2018]

Lenci, A. et al. (2000). SIMPLE: A general framework for the development of multilingual lexicon. *International Journal of Lexicography*, 13 (4), 249-263.

Leary, R., Vandenberghe, W. & Zeleznikow, J. (2004). Towards a financial fraud ontology: a legal modelling approach, *ICAIL 2003 Workshop on Legal Ontologies & Web based legal information management*, 1-33.

Masolo, C. et al. (2003). WonderWeb Deliverable D18: Ontology Library. *Laboratory for Applied Ontology*, ISTC-CNR.

McCarthy, D., Kilgarrif, A., Jakubíček, M. & Reddy, S. (2015). Semantic Word Sketches. *Corpus Linguistics (CL2015),* 1-5.

Miller, G. & Fellbaum, C. (2007). WordNet Then and Now. *Language Resources and Evaluation* 41(2), 209-214. Available at https://wordnet.princeton.edu/ [last accessed 17 May 2019]

Niles, I. & Pease, A. (2001). Towards a standard Upper Ontology.In *Proceedings of the Second International Conference on Formal Ontology in Information Systems*.Ogunquit. Available at http://www.adampease.org/professional/FOIS.pdf [last accessed 10 January 2019]

Pedersen, B.S. & Keson, B. (1999). SIMPLE–Semantic information for multifunctional plurilingual lexica: some examples of Danish concrete nouns. *Proceedings of the SIGLEX-99 Workshop*. Maryland. Available at http://clair.eecs.umich.edu/aan/paper.php?paper_id=W99-0507#pdf [last accessed 15 January 2019]

Periñán-Pascual, C. (2012). En defensa del procesamiento del lenguaje natural fundamentado en la lingüística teórica. *Onomázein*, 26, 13-48.

Periñán-Pascual, C. (2013). A knowledge-engineering approach to the cognitive categorization of lexical meaning. *VIAL - Vigo International Journal of Applied Linguistics*, 10, 85-104

Periñán-Pascual, C. & Arcas-Túnez, F. (2004). Meaning postulates in a lexico-conceptual knowledge base. *15th International Workshop on Databases and Expert Systems Applications, IEEE*, Los Alamitos (California), 38-42.

Periñán-Pascual, C. & Arcas-Túnez, F. (2005). Microconceptual-Knowledge Spreading in FunGramKB. *Proceedings of the 9th IASTED International Conference on Artificial Intelligence and Soft Computing*. Anaheim-Calgary-Zurich: ACTA Press, 239-244.

Periñán-Pascual, C. & Arcas-Túnez, F. (2010a). The architecture of FunGramKB. *Proceedings of the 7th International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA)*, 2667-2674.

Periñán-Pascual, C. & Arcas-Túnez, F. (2010b). Ontological commitments in FunGramKB. *Procesamiento del Lenguaje Natural*, 44, 27-34.

Periñán-Pascual, C. & Mairal-Usón, R. (2009). Bringing Role and Reference Grammar to Natural Language Understanding. *Procesamiento del Lenguaje Natural*, 43, 265-273.

Periñán-Pascual, C. & Mairal-Usón, R. (2010). La gramática de COREL: un lenguaje de representación conceptual. *Onomázein*, 21, 11-45.

Periñán-Pascual, C. & Mairal-Usón, R. (2011). The COHERENT Methodology in FunGramKB. *Onomázein*, 24,13-33.

Ruiz-de-Mendoza Ibáñez, F. & Mairal-Usón, R. (2009). Constructing meaning: a brief overview of the Lexical Constructional Model. In Mario Brdar (Ed.), *Converging and diverging tendencies in Cognitive Linguistics*. Amsterdam/Philadelphia: John Benjamins.

Ruppenhofer, J., Boas, H., & Baker, C. (2017). FrameNet. In P. Fuertes-Olivera (Ed.), *The Routledge Handbook of Lexicography*, (pp. 383-398). New York: Routledge.

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. *Proceedings of the 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing*, *RASLAN*, 2, 6-9.

Sartor, G., Casanovas, P., Biasotti, M.A. & Fernández-Barrera, M. (Eds.), (2011). *Approaches to legal ontologies, theories, domains, methodologies*, Berlin: Springer.

Thorgren, S. (2005). Transaction Verbs: A Lexical and Semantic Analysis of *Rob* and *Steal*. *Reports from the Department of Language and Culture*, 3, 1-44.

Valente, A. (2005). Types and roles of legal ontologies. In R. Benjamins, P. Casonovas, J. Breuker & A. Gangemi (Eds.), *Law and the semantic web* (pp. 65-76). Berlin: Springer.

Van Valin, R. (2005). *Exploring the Syntax-Semantics Interface*. Cambridge: Cambridge University Press.

Velardi, P., Pazienza, M., & Fasolo, M. (1991). How to Encode Semantic Knowledge: A Method for Meaning Representation and Computer-aided Acquisition. *Computational Linguistics*, 17(2), 153-170.