

Who stole what from whom?

A corpus-based, cross-linguistic study of English and Spanish verbs of stealing

Nicolás José Fernández-Martínez and Pamela Faber

University of Granada (Spain)

Drawing on the Lexical Grammar Model, Frame Semantics and Corpus Pattern Analysis, we analyze and contrast verbs of stealing in English and Spanish from a lexico-semantic perspective. This involves looking at the lexical collocates and their corresponding semantic categories that fill the argument slots of verbs of stealing. Our corpus search is performed with the Word Sketch tool on Sketch Engine. To the best of our knowledge, no study has yet taken advantage of the Word Sketch tool in the study of the selection preferences of verbs of stealing, let alone a semantic, cross-linguistic study of those verbs. Our findings reveal that English and Spanish verbs of stealing map out the same underlying semantic space. This shared conceptual layer can thus be incorporated into an ontology based on deep semantics, which could in turn enhance NLP tasks such as word sense disambiguation, machine translation, semantic tagging, and semantic parsing.

Keywords: verbs of stealing, cross-linguistic study, Lexical Grammar Model, semantic space, English/Spanish

1. Introduction

This paper presents a semantic, corpus-based study of English and Spanish verbs of stealing (e.g. *steal* ‘robar’, *rob* ‘atracar’, *embezzle* ‘desfalcar’, *plunder* ‘saquear’) that specifies the semantic categories for arguments and establishes cross-linguistic correspondence at both semantic and syntactic levels. Our analysis is based on the Lexical Grammar Model (Faber and Mairal, 1999), which organizes the verbal lexicon into semantic domains, and also incorporates insights from Frame Semantics (Fillmore, 1982, 1985; Fillmore and Baker, 2010; Ruppenhofer *et al.*, 2017).

Using Corpus Pattern Analysis (Hanks, 2004, 2012, 2013; Hanks and Jezek, 2010), we perform a semi-automatic corpus search for the lexical collocates and

semantic categories that populate the argument slots of English and Spanish verbs of stealing, organized according to their logDice salience score (Rychlý, 2008). The basic inventory of semantic categories is initially derived from WordNet lexical hypernyms (Miller and Fellbaum, 2007) and then adapted for our purposes. Our findings suggest that verbs of stealing in English and Spanish map out the same semantic space. This makes it possible to specify a shared conceptual scenario of theft and robbery lexicalized in both languages. In addition, this type of corpus-based, cross-linguistic studies has potential implications for deep semantic frameworks in the conceptual modeling of ontologies (Periñán and Arcas, 2007; Velardi *et al.*, 1991) and other NLP tasks such as word sense disambiguation, machine translation, semantic tagging, and semantic parsing.

The remainder of this paper is organized as follows. Section 2 introduces the Lexical Grammar Model as well as Frame Semantics. Section 3 describes the methodology and explains how the verbs were compared in both languages. Section 4 gives the results obtained and discusses their implications. Finally, Section 5 presents the conclusions that can be derived from this research.

2. The lexico-semantic dimensions of verbs of stealing

2.1 Lexical grammar model

According to the Lexical Grammar Model (LGM) (Faber and Mairal, 1999, 2017), the lexicon is the interface between semantics and syntax, and semantic organization is regarded as a predictor of syntactic behavior. In this sense, the LGM highlights the interaction between the microstructural and macrostructural information encoded in lexical units (Butler, 2003). The verbal lexicon is organized in domains and subdomains based on the factorization of shared meaning components using Stepwise Lexical Decomposition (Dik, 1978). The position of each verb in the hierarchy is based on its syntagmatic and paradigmatic information as well as its interaction with the semantics-syntax interface. In this way, lexical domains provide a structured inventory of semantically related lexemes headed by a hypernym, in terms of which all of the other lexical units are defined (Faber and Mairal, 1998a, 1999). The prototypical cognitive-conceptual schema of verbs of possession involves an animate possessor and a possessed entity (Faber and Mairal, 1998b; Harley, 2003; Heine, 1997):

- (1) *John has a car.*

Verbs of possession can also reflect different types of ownership as well as transfer or loss of ownership (Faber and Mairal, 1998c). The subdomains of this lexical

field activate more specific conceptual-cognitive schemas (Boas, 2013; Faber and Mairal, 2017). Table 1 shows a segment of the domain of POSSESSION, specifically from the subdomain *To come to have something*.

Table 1. Subdomain from the LGM domain of POSSESSION

<i>To come to have something</i>	
find:	to come to have something that one has been looking, usually after searching for it.
get:	to come to have something as a result of receiving, earning, buying, etc.
land:	to get something (something difficult to get that others also want [informal]).
secure:	to get something after making a great effort (formal).
obtain:	to get something (formal).
procure:	to obtain something difficult to get with care (formal).
acquire:	to obtain something with effort, adding it to previous possessions.

This subdomain encodes the inception of a possessive event. The superordinate terms *find* and *get* feature the most prominent semantic properties and also have a greater number of complementation patterns than their subordinates. In this sense, the superordinate determines domain membership, and its semantics is inherited by the more specific verbs in the lexical hierarchy as reflected in the semantic nucleus of their meaning definitions. At the same time, subordinate terms specify their own differentiating parameters.

As shown in Table 1, the hyponyms of *get* are *land*, *secure*, and *obtain*. *Procure* and *acquire* are inherited from and defined in terms of *obtain*. The *differentiae* specify the differentiating parameters specific to each verb, such as its level of formality, the difficulty or effort invested in getting something, or any other temporal or logical relation related to inceptive possession. The more subordinate members of the hierarchy have more specific meanings as well as fewer complementation patterns.

The LGM can also be used to juxtapose and compare lexical domains in different languages. Table 2 shows the subdomain of *To take something away from somebody without the right to do so* from the lexical domain of POSSESSION, which refers to theft and robbery verbs in English and the corresponding domain in Spanish. Lexical gaps in either of the languages are indicated in italics by paraphrases in the case of verbal lexemes with specific differentiating parameters (e.g. *to steal something small and of little value* for *hurtar*) or their superordinate terms (e.g. *saquear* for *despoil*). Phrasal verbs (*clean somebody/something out* and *make off with something*) also appear in italics.

Nonetheless, for a more fine-grained analysis, these verbs should not only be considered individually, but also in a wider context. This can be accomplished by incorporating certain premises and insights from Frame Semantics.

Table 2. Subdomain of verbs of stealing from LGM domain of POSSESSION

Verbs of possession	Verbos de posesión
<i>To take something away from somebody without the right to do so</i>	<i>Quitar algo de valor a alguien</i>
steal: to take something away from somebody without their permission and not intending to return it (unlawfully).	robar: quitar algo de valor a alguien con violencia o engaño (delito).
rob: to steal something (especially money/property) from somebody/institution.	atracar: robar en un sitio/a alguien con amenazas/usando la fuerza.
<i>to hold up</i>	asaltar: atracar algo o alguien irrumpiendo violentamente.
<i>to stick up</i>	robar
thieve: to steal (old fashioned).	desfalcar: robar alguien dinero puesto en su cuidado.
embezzle: to steal money placed in your care for your own purposes.	descantillar: desfalcar (arcaico).
<i>to steal something small and of little value</i>	hurtar: robar poco dinero u objetos de poco valor sin violencia.
<i>to steal something by keeping a small quantity of it for yourself (when you should give it to somebody else).</i>	sisar: hurtar algo, especialmente en la compra diaria.
purloin: to steal something especially small (literary) (formal).	sustraer: robar algo sin violencia.
shoplift: to steal things from shops by taking them from the shelves and hiding them under clothes or in a bag.	<i>hurtar en tiendas</i>
pilfer: to steal things that are small/of little value especially continuously over a period of time.	ratear: robar cosas de poco valor a lo largo de un tiempo.
pinch: to steal something especially small directly off somebody (BrE) (informal).	mangar: robar algo sin gran valor (informal).
<i>steal</i>	pisar: robar algo, cogiéndolo antes que otra persona que también lo quería.
filch: to steal things that are small/of little value in a very secretive way (informal).	birlar: robar algo a alguien con engaño o habilidad (coloquial).
lift: to steal (informal).	soplar: robar algo a alguien engañándole (coloquial).
nick: to steal (BrE) (informal).	afanar: robar algo a alguien con habilidad.
swipe: to steal something by removing it quickly.	robar
rustle: to steal cattle/horses (AmE).	<i>robar ganado (vacas/caballos)</i>
<i>to clean somebody/something out</i>	limpiar: robar a alguien, especialmente quitándoselo todo en el juego.
<i>steal</i>	desvalijar: robar todo lo que alguien lleva encima/todo lo que hay en un lugar.

Table 2. (continued)

Verbs of possession	Verbos de posesión
<i>to make off with something</i>	arramblar : robar algo, llevándolo de un sitio todo lo que hay con abuso y codicia.
plunder : to take things violently from a place in time of war or disorder.	saquear : robar las posesiones de los vencidos, apoderándose de lo que encuentran en el lugar.
pillage : to plunder (archaic).	pillar saquear (arcaico).
despoil : to plunder (formal) (literary).	<i>saquear</i>
loot : to take things in large quantities from buildings (shops, churches, houses), causing damage, during a violent event (riot/battle) or after a natural catastrophe (hurricane/typhoon).	<i>saquear</i>

2.2 Frame semantics

Frame Semantics links linguistic forms to cognitive structures or schemas (i.e. frames) that shape the way that we conceive and interact with everything related to human experience and world knowledge (Barsalou, 1992; Fillmore, 1982, 1985; Fillmore and Baker, 2010; Ruppenhofer *et al.*, 2017). The linguistic forms that encode these events and processes are signals that evoke and activate these frames in our minds, and allow us to understand what is taking place.

FrameNet is the computational implementation of Frame Semantics (Fillmore *et al.*, 1998; Ruppenhofer *et al.*, 2010). It is a corpus-driven, lexico-semantic database that captures semantic frames and their corresponding lexicalization in English (Boas, 2005). There are also FrameNets for Spanish (Subirats, 2009) and many other languages.

In the English FrameNet and Spanish FrameNet, verbs of possession invoke frame structures such as Possession, Giving, Taking, Getting, Commercial transaction, Theft and Robbery, each of which activates its own frame elements. The lexical collocates that can potentially fill in those attributes are termed ‘values’ in the sense of Barsalou (1992:30–31). For example, *money* would be a possible filler or value for GOODS. These values are, however, not represented in FrameNet.

Verbs of stealing embedded in the Theft and Robbery frames activate similar cognitive schemas (Boas, 2013:127). At the same time, they ‘profile’ frame elements in different ways (Langacker, 2007:438–441). This means that despite sharing a certain core of background knowledge or ‘background frame’ (Goldberg, 1995: 45–48, 2010: 41), these verbs may put greater or lesser emphasis upon different core elements such as the goods or the victim. For example, *steal* profiles the

Table 3. Frame structures of verbs of possession

Verbs of possession	Frame structure	Core frame elements
<i>have, own, belong...</i>	Possession	OWNER, POSSESSION
<i>receive, get, obtain, acquire...</i>	Getting	RECIPIENT, THEME _{PHYSICAL OBJECT}
<i>give, hand, pass, provide...</i>	Giving	DONOR, RECIPIENT, THEME _{PHYSICAL OBJECT}
<i>steal, pilfer, filch...</i>	Theft	PERPETRATOR _{SENTIENT} , GOODS, SOURCE _{SOURCE} , VICTIM _{SENTIENT}
<i>rob, hold up...</i>	Robbery	PERPETRATOR _{SENTIENT} , SOURCE _{SOURCE} , VICTIM _{SENTIENT}
<i>grab, take...</i>	Taking	AGENT _{SENTIENT} , THEME _{PHYSICAL OBJECT} , SOURCE _{SOURCE}

perpetrator and the goods, whereas *rob* profiles the perpetrator and the victim (Dux, 2018; Thorgren, 2005). This difference in profiling character is semantically motivated and ultimately reflected in the syntax of their arguments. The profiled elements are obligatory whereas others are optional.

One of the main issues in a frame-semantics model such as FrameNet is its lack of specificity and granularity in frame structures. Frame structures cover a large number of lexical units, but do not provide an in-depth characterization of the semantic types that populate each frame element. Although concerns of this type have been voiced (Boas, 2005: 474–475, 2008, 2013), this issue has not as yet been satisfactorily addressed despite attempts to unify other frame databases and convert them into a single cross-lingual one (Gilardi and Baker, 2018; Gruzitis and Dannélls, 2017).

Our study differs from other frame-based cross-linguistic studies of verbs of stealing (Dux, 2011, 2018) in that we focus on semantic frames as a framework for semantic valency. For this purpose, we analyze the semantic categories that appear in the frame elements or argument slots of verbs of stealing and specify their selection preferences.

3. Data and method

Verbs of stealing belong to the subdomain *To take something away from somebody without the right to do so* within the lexical domain of POSSESSION. Our research analyzes the complementation patterns as well as the syntagmatic and paradigmatic information of these verbs. For this purpose, we use Corpus Pattern Analysis (Hanks, 2004, 2012, 2013; Hanks and Jezek, 2010) to study their valency and

selection preferences, based on the lexical collocates and the corresponding semantic categories that fill their argument slots. Section 3.1 presents the corpora that we analyze in the search of lexical collocates. Section 3.2 describes the steps and processes in obtaining lexical collocates and their organization in semantic categories. Finally, Section 3.3 covers the flowchart used to establish cross-linguistic correspondence.

3.1 Corpora

For our semi-automatic corpus study, we use Sketch Engine, an online web platform that has multilingual corpora and user-friendly tools (Kilgarriff *et al.*, 2014), and which provides data related to word associations, usage patterns, and language change. One of these 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: 372). This tool provides us with the most frequent lexical collocates for each syntactic constituent.

Verbs of stealing and their lexical collocates are extracted from different monolingual corpora, namely, the BNC corpus (2007) and the English Web 2013 and 2015 corpora (enTenTen13 and enTenTen15) (Jakubíček *et al.*, 2013). The BNC corpus stores 100 million words of written and spoken texts in British English from the second half of the 20th century. The enTenTen13 and enTenTen15 corpora are composed of texts obtained from the Web and have around 15 billion words from every geographical variety of English. Spanish data are obtained from the Spanish Web 2011 corpus (esTenTen11) (Kilgarriff and Renau, 2013) and the Spanish Timestamped 2014–2018 corpus (Trampus and Novak, 2012). The esTenTen11 corpus contains 9.5 billion words from European and American Spanish. The Spanish Timestamped 2014–2018 corpus stores 6 billion words of European and American Spanish news articles.

3.2 Procedure for the extraction of lexical collocates and grouping into semantic categories

For the corpus search, we query each verb (21 English stealing verbs and 19 Spanish stealing verbs) in the Word Sketch tool, and select the gramrels “Subject of X”, “Object of X” and “Prepositional Phrases” to look for the obligatory and/or optional syntactic arguments, linked to the semantic valency or frame structure of each verb.

This generates tables with each syntactic constituent and its most frequent lexical collocates. Accordingly, for verbs such as *steal*, *pilfer* or *filch*, we take into

Select gramrels:
☐ All

<input type="checkbox"/> inimitive objects or X	<input type="checkbox"/> It's X to ...	<input type="checkbox"/> modifiers or X	<input type="checkbox"/> nouns and veros modified by X
<input type="checkbox"/> objects of "X"	<input checked="" type="checkbox"/> objects of X	<input type="checkbox"/> particles after X	<input type="checkbox"/> particles after X with object
<input type="checkbox"/> possessors of X	<input checked="" type="checkbox"/> prepositional phrases	<input type="checkbox"/> pronominal objects of X	<input type="checkbox"/> pronominal possessors of X
<input type="checkbox"/> pronominal subjects of X	<input checked="" type="checkbox"/> subjects of X	<input type="checkbox"/> subjects of "be X"	<input type="checkbox"/> verbs before X
<input type="checkbox"/> verbs before X and noun	<input type="checkbox"/> verbs with X as object	<input type="checkbox"/> verbs with X as subject	<input type="checkbox"/> verbs with particle "and" and X as object

Figure 1. Gramrel selection for corpus search

account not only the subject and direct object arguments, but also any nearby prepositional phrase (PP), since this phrase, which encodes the source or victim affected by the act of stealing, functions as an oblique object (Berman, 1982). This argument is commonly found in English theft verbs (e.g. *steal something from someone/from a place*) (Levin, 1993:128–129).

However, in robbery verbs such as *rob*, *hold up*, and *stick up*, the subject and direct object arguments usually represent the perpetrator and victim/source, respectively. As can be observed in *rob*, the goods may sometimes be syntactically realized by an optional PP introduced by *of*:

- (2) Three off-duty-soldiers were robbed *of their cellphones and wristwatches* [...].
(enTenTen15)

Despite the fact that these syntactic arguments in verbs of stealing are generally optional, they are semantically relevant.

Some English verbs of stealing also have more than one syntactic pattern. This is the case of *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*).

In contrast, many Spanish verbs of stealing are generally ditransitive and their syntactic behavior resembles that of verbs of giving (Enghels and Whylin, 2015:113–114) as reflected in the double object construction. Since they are by default ditransitive (e.g. *sustraer algo a alguien* ‘purloin something from somebody’ or *robar algo a alguien* ‘steal something from somebody’), special attention is paid to their direct object argument (i.e. the goods frame) and to their indirect object in the form of a PP activating the victim frame. We also consider any optional PP (*de* ‘from’ or *en* ‘in’), which might potentially refer to the source frame.

A few Spanish verbs of stealing (i.e. robbery verbs), such as *atracar* ‘rob’, are monotransitive, which means that only the subject and direct object are considered. These grammatical roles generally refer to the perpetrator and victim or source frames, respectively. At times, Spanish verbs of stealing can have more than one syntactic pattern. For instance, *robar* ‘steal’ participates in two syntactic constructions, one ditransitive and another monotransitive: *robar la cartera a una persona* ‘steal somebody’s wallet’ or *robar un banco* ‘rob a bank’. The same applies to verbs such as *saquear* (*saquear algo a alguien* ‘plunder something from some-

↔ ☰ 🔍 ✕	↔ ☰ 🔍 ✕	↔ ☰ 🔍 ✕
subjects of "steal"	objects of "steal"	prepositional phrases
thief 9.91 ...	show 8.25 ... stole the show	"steal" from ... 5.23% ...
hacker 8.64 ...	bike 7.6 ... bike stolen	"steal" by ... 1.06% ...
robber 7.18 ... robbers stole	heart 7.45 ...	"steal" in ... 0.93% ...
criminal 7.15 ...	car 7.34 ...	"steal" of ... 0.33% ...
burglar 6.87 ...	election 7.26 ...	"steal" during 0.2% ...
suspect 6.67 ... suspect stole	spotlight 7.22 ... stole the spotlight	...
nazis 6.63 ... stolen by the Nazis	money 7.2 ...	"steal" on ... 0.19% ...
someone 6.51 ... someone stole	glance 7.18 ...	"steal" per ... 0.17% ...
grinch 6.4 ... Grinch who stole	land 7.13 ...	"steal" at ... 0.14% ...
gang 6.32 ...	thunder 7.13 ...	"steal" for ... 0.14% ...
	identity 7.06 ...	"steal" over ... 0.07% ...
		"steal" into ... 0.07% ...

Figure 2. Lexical collocates of the arguments of steal

body', *saquear un lugar* 'plunder a place', *saquear algo de un lugar* 'plunder something from a place') or *desfalcar* (*desfalcar algo a alguien* 'embezzle something from somebody', *desfalcar algo de un lugar* 'embezzle something from a place').

The next step focuses upon the extraction of lexical collocates for each syntactic slot. We annotate those lexical collocates that are the most frequent and thus the most representative for each semantic category. For the semantic grouping of lexical collocates, we use an adapted inventory of WordNet hypernyms (Miller and Fellbaum, 2007) to avoid multiplicity and overlapping. When necessary, these semantic categories are made more specific. Table 4 shows an alphabetical list of these semantic categories together with instances obtained from our corpus search.

Because of the variation in corpus size (Rychlý, 2008), frequency is measured in terms of the logDice salience score, situated next to each lexical collocate. We then calculate the means of logDice salience scores obtained from different corpora. For instance, the goods *car* in the object list of collocates of *steal* has a score of 8.96 in the BNC corpus, a score of 7.25 in the enTenTen13 corpus and a score of 7.34 in the enTenTen15 corpus, giving an average of 7.85. This is the value in our

↔	☰	🔍	✕	↔	☰	🔍	✕	↔	☰	🔍	✕
subjects of "robar"				objects of "robar"				prepositional phrases			
ladrón	9.5	...		vehículo	8.75	...		"robar" a ...	6.76%	...	
delincuente	9.02	...		vehículos robados				"robar" en ...	4.61%	...	
delincuentes robaron				auto	8.59	...		"robar" de ...	2.24%	...	
desconocido	7.08	...		autos robados				"robar" por ...	1.25%	...	
desconocidos robaron				dinero	8.32	...		"robar" con ...	0.7%	...	
sujeto	7.02	...		moto	8.21	...		"robar" hasta	0.23%	...	
asaltante	6.9	...		balón	8.04			
asaltantes robaron				plata	7.96	...		"robar" desde	0.18%	...	
estan	6.49	...		cartera	7.93			
QUE NOS ESTAN ROBANDO				camioneta	7.84	...		"robar" durante	0.17%	...	
políticos	6.47	...		pertenencia	7.76			
los políticos roban				carro	7.46	...					
dinero	6.32	...									
banda	6.27	...									

Figure 3. Lexical collocates of the arguments of robar

tables. As a final step, we organize the lexical collocates of English and Spanish verbs of stealing in semantic categories and establish cross-linguistic correspondence between them based on semantic and syntactic similarity.

This procedure, however, has certain limitations. On the one hand, the Word Sketch tool sometimes retrieves erroneous lexical collocates, duplicates or corpus junk (Kilgarriff *et al.*, 2010:378), probably because of a malfunction in the syntactic parser and part-of-speech tagger. Since our concern is with the basic meaning of stealing, secondary or metaphorical meanings (e.g. *steal a glance/show*) are disregarded. Irrelevant or erroneous collocates are manually discarded. At times, there are only erroneous lexical collocates for certain syntactic constituents in some stealing verbs. For instance, the subject list of the lexical collocates of *shoplift* provides collocates such as *felony*, *misdemeanor*, *again*, or *completely*. Other word sketches display very few lexical collocates (or even none at all) in the case of archaic and old-fashioned verbs such as *pillar* ‘pillage’ or *pisar* ‘steal’, or in the case of non-prototypical stealing verbs such as *mangar* ‘pinch’. In those cases, it is necessary to manually search concordances and annotate possible lexical collocates without any mention of their logDice salience score. When no concordances are found in the corpora, we use dictionaries such as *Collins*, *Cambridge*,

Table 4. Adapted WordNet's hypernyms for semantic classes

Semantic categories	Lexical collocates
ANIMAL	cat, dog, horse, fox, cattle, livestock...
ART	artwork, painting, sculpture...
ASSETS	money, cash, riches, wealth, pounds, dollars, euros, land, estate, property, booty, treasure...
CASE	wallet, purse, bag, wristband...
CONTAINER	bottle, box...
COSMETICS	lipstick, nail polish, razor, perfume...
DOCUMENT	book, passport, ticket, card...
DRUG	medicament, vitamins, minerals, booze, cigarette, beer, wine, weed, heroin...
ELECTRONICS	TV, phone, radio, laptop, stereo...
FOOD	meat, fruit, vegetables, grains, legumes, milk, water...
FUEL	gas, petrol...
GARMENT	clothes, t-shirt, underwear, skirt, trousers, handkerchief...
HUMAN	thief, criminal, tourist, taxpayer, woman, baby, gang, band, mob, government, state, company, multinational...
METAL	gold, silver...
MINERAL	diamond, sapphire, ruby...
ORNAMENT	jewelry, gold, necklace...
PLACE	shop, supermarket, church, bank, garden, countryside, farm, house...
TOOL	hammer, drill, drumstick...
VEHICLE	car, bike, auto, caravan...
WEAPON	gun, firearm, rifle...

Oxford, Longman, Diccionario de la Real Academia Española, WordReference or *Diccionario Salamanca de la Lengua Española* to obtain definitions and examples of lexical collocates. As a last resort, certain combinations of verb plus collocates are obtained from the Google search engine.

3.3 Cross-linguistic equivalence

To establish cross-linguistic equivalence between English and Spanish verbs of stealing, we apply a modified version of the flowchart in Baisa *et al.* (2016), which has been adapted to establish cross-linguistic equivalence in terms of frame semantic valency (see Figure 4). Syntactic valency is only taken into account to distinguish among different shades of matching correspondence.

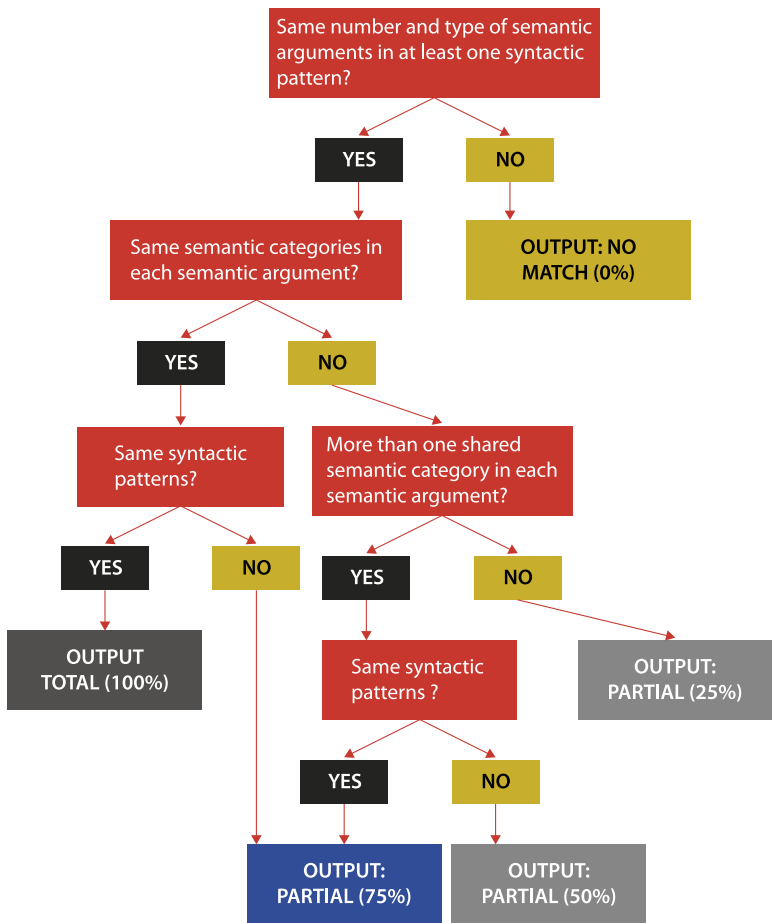


Figure 4. Flowchart of process for establishing cross-linguistic correspondence

4. Results and discussion

This section presents and discusses the findings of our corpus study. Section 4.1 describes the inventory of semantic categories found in each argument slot for each language and the more abstract semantic categories in the conceptual scenario of theft and robbery. Section 4.2 analyzes the matching rate of cross-linguistic correspondence between English and Spanish verbs of stealing.

4.1 Semantic categories in verbs of stealing

Table 5 and Table 6 give the full inventory of semantic categories for English and Spanish verbs of stealing.

Table 5. Semantic categories of English verbs of stealing

Semantic categories of English verbs of stealing					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	95	ASSETS	19	PLACE	55
ANIMAL	5	ANIMAL	14	HUMAN	43
		DRUGS	10	ANIMAL	2
		CASE	10		
		FOOD	9		
		ELECTRONICS	8		
		ORNAMENT	6		
		GARMENT	5		
		VEHICLE	4		
		DOCUMENT	3		
		FUEL	2		
		COSMETICS	2		
		TOOL	2		
		MINERALS	2		
		METAL	2		
		ART	2		

Table 6. Semantic categories of Spanish verbs of stealing

Semantic categories of Spanish verbs of stealing					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	98	ASSETS	44	PLACE	52
ANIMAL	2	CASE	10	HUMAN	48
		VEHICLE	9		
		ELECTRONICS	7		
		ANIMAL	5		
		ORNAMENT	5		
		FUEL	4		
		GARMENT	4		
		METAL	2		
		FOOD	2		
		MINERAL	2		
		HUMAN	2		
		ART	2		
		WEAPON	1		

As can be observed, the agent that steals is always an animate entity, whether human (an individual or group of individuals) or an animal. Despite the generic character of the HUMAN category, there are humans that are far less likely to engage in acts of stealing. Indeed, at the axiological level, stealing verbs have very negative connotations because of the encyclopedic knowledge associated with the act of stealing and its typical agents. In this sense, perpetrators of stealing are negatively viewed (e.g. *thieves, bandits, robbers, burglars, pickpockets...*), whereas other more positively evaluated social groups (e.g. *nuns, volunteer workers, or one-year-old babies*) would be highly unlikely participants in such illicit activities (Faber and Mairal, 1999:97). Furthermore, each stealing verb has its typical set of perpetrators, depending on their semantics. For example, a one-year old baby could not feasibly be involved in embezzling activities, nor would the baby be able to rob a bank. This means that *embezzle* requires the perpetrator to work for a company or an organization and be responsible for the goods owned by that company or organization. A small child or a person not employed by the company or organization could thus not assume this role. In other stealing verbs such as *plunder, loot, pillage* and *despoil*, the human agent is also different from those of the other verbs in the hierarchy. The typical agents of these verbs are usually large groups of people who, in times of war and violence, take advantage of the turmoil to unlawfully take large quantities of precious goods. In sum, even though a large percentage of stealing agency involves humans, not all humans are likely to steal. Nevertheless, for the sake of simplicity, we choose to represent the HUMAN category in its most generic sense. This means that the lexicosemantic characteristics associated with the human participants in specific types of stealing can be derived from our encyclopedic knowledge regarding each verb and the context evoked.

In contrast to the small number of semantic categories of perpetrators, there is a much wider range of goods that can be stolen. In both languages, the goods most frequently taken are ASSETS – though this is even more frequent in Spanish. The most significant difference is that in English the victim can be ANIMAL. Oddly enough, this category does not appear in Spanish verbs, despite the fact that food can certainly be stolen from animals in Spain and Latin America. Most of the semantic categories listed can be regrouped into more general semantic categories such as ARTEFACT, ANIMATE ENTITY, and NATURAL OBJECT without losing granularity.

Table 7 shows the general semantic categories that participate in the conceptual scenario of theft and robbery.

ARTEFACT is anything that is man-made or which has somehow been modified or altered by humans, such as ASSETS, ELECTRONICS, VEHICLE, ORNAMENT, GARMENT, TOOL, or ART. NATURAL OBJECT refers to anything that has not been modified by human intervention (i.e. FOOD, METAL, MINERAL, or FUEL). Certain

Table 7. Semantic categories of conceptual scenario of theft and robbery

Semantic categories in the conceptual scenario of theft and robbery		
PERPETRATOR	GOODS	VICTIM/SOURCE
Semantic categories	Semantic categories	Semantic categories
ANIMATE ENTITY	ARTEFACT	ANIMATE ENTITY
	NATURAL OBJECT	PLACE
	ANIMATE ENTITY	

categories such as DRUG or FUEL can be subsumed under ARTEFACT or NATURAL OBJECT, depending upon the lexical collocate.

Petrol, for instance, is obtained from *petroleum* through human intervention, whereas *wood*, another type of FUEL, does not necessarily require human intervention. Some medicinal drugs such as herbal roots can be categorized as NATURAL OBJECT. In contrast, *paracetamol* is an artificial man-made drug sold at a pharmacy, that is, an ARTEFACT. The conceptual scenario of theft and robbery is governed by these general semantic categories. Each verb of stealing thus activates this scenario, though each has its own characteristics that differentiate it from the other verbs in the domain. For instance, *rustle* (Tables 8 and 9) refers to the stealing of a member of the semantic category ANIMAL, especially cattle bred on a farm.

Table 8. Lexical collocates of *rustle*

<i>Rustle something from someone/a place</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
bandit	0.73	cattle	6.30	rancher	2.22
		herd	1.81	ranch	1.27
		sheep	1.81	herd	1.00
		livestock	1.24	employee	0.14
		horse	1.04		
		cow	1.00		

Table 9. Semantic categories of *rustle*

<i>Rustle something from someone/a place</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	100	ANIMAL	100	HUMAN	51
				PLACE	49

The object of *plunder* and its Spanish correspondence *saquear* (Tables 10, 11, 12 and 13) tends to be ASSETS since what is taken is generally something that is valuable. That is why other semantic categories such as FUEL, METAL, or MINERAL (i.e. natural resources or substances) are also quite commonly found with such verbs, together with many others typically taken in violent events (e.g. WEAPON, FOOD, ORNAMENT...).

Table 10. Lexical collocates of *plunder*
Plunder something from someone/a place / plunder a place

PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
invader	6.67	booty	6.52	<i>treasury</i>	6.77
pirate	4.25	treasure	5.88	<i>tomb</i>	5.95
looter	4.24	riches	3.95	<i>caravan</i>	5.92
conquistador	4.07	wealth	3.77	<i>coffer</i>	4.17
imperialist	3.90	mineral	2.56	<i>Egyptian</i>	3.88
Nazis	3.79	gold	2.38	<i>taxpayer</i>	3.85
multinational	2.89	diamond	1.11	victim	3.59
				farmhouse	3.56
				<i>monastery</i>	3.49
				<i>Congo</i>	3.14
				backyard	2.66
				<i>inhabitant</i>	2.55
				<i>merchant</i>	1.70
				land	0.71
				church	0.63

Note: Collocates in italics are realized as direct objects, contrasting with the remaining ones that typically appear in optional adjuncts introduced by *from*.

Table 11. Semantic categories of *plunder*
Plunder something from someone/a place / plunder a place

PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	100	ASSETS	54	PLACE	54
		METAL	26	HUMAN	46
		MINERAL	20		

Table 12. Lexical collocates of *saquear*

<i>Saquear algo a alguien/de un lugar / saquear un lugar</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
turba	8.51	riqueza	7.14	<i>arcas</i>	9.59
ladrón	7.64	petróleo	4.93	<i>supermercado</i>	8.66
vándalo	7.42	oro	3.79	<i>erario</i>	8.11
encapuchados	6.77	madera	3.34	<i>tienda</i>	7.69
multinacionales	2.79	dinero	3.22	país	7.15
		mineral	2.22	<i>tumba</i>	6.74
		botín	2.19	<i>bolsillo</i>	6.62
		cobre	1.87	<i>casa</i>	5.85
				pueblo	4.41
				jubilados	1.96

Note: Collocates in italics are realized as direct objects, contrasting with the remaining ones that typically appear in optional adjuncts introduced by *de*.

Table 13. Semantic categories of *saquear*

<i>Saquear algo a alguien/de un lugar / saquear un lugar</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	100	ASSETS	34	HUMAN	64
		FUEL	30	PLACE	36
		METAL	20		
		MINERAL	16		

Shoplift (Tables 14 and 15) involves stealing any type of goods typically sold in stores such as DRUGS, TOOL, FOOD, ELECTRONICS, COSMETICS, or GARMENT. As expected, no examples referring to the victim frame are found. It is worth mentioning that in our corpus search the number of concordances referring to female shoplifters is greater than that of male shoplifters. This reinforces the claim that shoplifting is a predominantly female criminal activity, according to quantitative sociological and criminological studies (Marsh *et al.*, 2006:142).

Table 14. Lexical collocates of *shoplift*

<i>Shoplift something from a place</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
female	N/A	underwear	1.27	Walmart	4.69
people	N/A	medicament	1.02	counter	3.41
teens	N/A	groceries	0.92	supermarket	3.02
women	N/A	condom	0.90	store	2.70
customer	N/A	iPod	0.89	Tesco	2.07
men	N/A	peaches	0.86	bookstore	1.54
daughter	N/A	booze	0.83	drugstore	1.33
kids	N/A	DVDs	0.76	shop	0.76
		cologne	0.76		
		lipstick	0.75		
		handkerchief	0.69		

Table 15. Semantic categories of *shoplift*

<i>Shoplift something from a place</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	100	DRUGS	18	PLACE	100
		TOOL	18		
		FOOD	18		
		ELECTRONICS	16		
		COSMETICS	15		
		GARMENT	15		

In addition, we are surprised to find examples that seem to contradict encyclopedic knowledge. More specifically, *embezzle* (Tables 16 and 17), which is typically associated with *money* (Faber and Mairal, 1999:97), is found to be related to many other goods, such as *Bible*, *paracetamol*, *reptile*, *heroin*, or *meat*. Here we reproduce a few corpus examples:

- (3) never medicate it by urself, could bcome worse go to the doctors if it have got worse and maybe *embezzle some paracetamol or some ibuprofen* to help with the niggle. (enTenTen13)
- (4) Di Pisa accused the Americans of defrauding him, while the La Barberas accused Di Pisa of *embezzling the missing heroin*. (enTenTen13)
- (5) The *embezzled Porsche* was recovered in the garage. (enTenTen13)

This disparity in the range of possible lexical collocates, which are not linked to *money*, can only be explained if embezzling is understood as illegally taking any type of goods owned by a company or organization, for which a company employee is responsible. Embezzling a bible can be a possible scenario if a librarian illegally takes it from the library. In the same way, a person who works for a pharmaceutical company can embezzle DRUGS (e.g. paracetamol), a person who works for a car company can embezzle cars, and a member of the mafia can embezzle money, weapons or drugs.

Table 16. Lexical collocates of *embezzle*

<i>Embezzle something from someone/a place</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
accountant	4.86	fund	5.89	employer	3.45
bookkeeper	4.85	money	3.49	treasury	2.98
dictator	4.37	property	0.93	bank	2.11
treasurer	2.99	Bible	0.85	company	1.95
businessman	1.52	material	0.82	firm	1.94
employee	1.21	paracetamol	0.67	mob	1.62
banker	0.84	selenium	0.55	estate	1.35
		sapphire	0.49	organization	1.23
		reptile	0.38	charity	1.16
		ibuprofen	0.38	union	1.01
		Porsche	0.35	Church	0.98
		heroin	0.17	client	0.84
		meat	0.04	business	0.76

Table 17. Semantic categories of *embezzle*

<i>Embezzle something from someone/a place</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	100	ASSETS	52	HUMAN	75
		DOCUMENT	16	PLACE	25
		MINERALS	9		
		DRUGS	8		
		ANIMAL	7		
		VEHICLE	7		
		FOOD	1		

4.2 Cross-linguistic correspondence

For cross-linguistic equivalence, it is first necessary to examine each pair of translation equivalents. This means analyzing shared semantic arguments, shared semantic categories, and finally syntactic valency. This section discusses the most frequent verbs of stealing, such as *steal* and *robar*, in terms of different corpus-related or lexical discrepancies.

Steal and *robar* (Tables 18, 19, 20 and 21) share the same semantic arguments in type and number in the syntactic constructions in which they participate when referring to a theft event. They also largely share the same semantic categories for each semantic argument. Though the lexical collocates and semantic categories do not exactly match (e.g. *steal* + GARMENT OR WEAPON), this can be accounted for in terms of a lower logDice salience score that makes those lexical collocates go unnoticed. In this respect, intuition plays a key role when corpus instances are missing. In such cases, a search in dictionaries or the Web often help to confirm or reject our assumptions. In short, their match is partial (75%) because of their different syntactic behavior.

Indeed, constructional patterns play a decisive role in sorting out cases of total or partial equivalence. For example, *robar* – which can correspond to both *steal* and *rob* in English – admits an additional construction in which the source is the direct object (e.g. *robar un banco*), a construction that is not found in English (e.g. **steal a bank*). Moreover, the default syntactic constructions of these verbs differ since *robar* is ditransitive, while *steal* can be both monotransitive and ditransitive (Berman, 1982).

Table 18. Lexical collocates of *steal*

Steal something from someone/a place					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
thief	10.08	car	7.85	car	8.71
burglar	7.49	bike	7.22	museum	8.22
robber	6.66	wallet	7.07	garage	7.92
someone	6.51	money	7.06	house	7.86
gang	6.48	jewellery	6.35	people	7.46
suspect	5.84	food	6.31	victim	7.19
attacker	5.52	land	6.25	employer	7.18
hacker	5.38	card	6.25	taxpayer	7.05
employee	4.84	horse	6.19	farm	6.80
fox	4.81	cattle	6.14	neighbor	6.71
criminal	4.67	passport	5.59	store	6.44

Table 18. (continued)

Steal something from someone/a place

PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
monkey	3.33	phone	5.41	purse	6.43
Google	1.90	radio	4.89	residence	4.88
country	1.59	laptop	4.59	car	8.71
		baby	4.56	museum	8.22

Table 19. Semantic categories of *steal**Steal something from someone/a place*

PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	63	VEHICLE	14	PLACE	50
ANIMAL	37	CASE	13	HUMAN	50
		ASSETS	12		
		ORNAMENT	11		
		FOOD	11		
		DOCUMENT	11		
		ANIMAL	11		
		ELECTRONICS	9		
		HUMAN	8		

Table 20. Lexical collocates of *robar**Robar algo a alguien/en o de un lugar / robar un lugar*

PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
delincuente	9.65	vehículo	8.96	vivienda	8.15
ladrón	9.57	dinero	8.41	rico	7.82
sujeto	7.75	cartera	8.14	supermercado	7.65
desconocido	7.70	teléfono	7.43	tienda	7.56
banda	7.45	joya	7.32	pasajero	7.46
hackers	7.45	bolso	7.10	anciano	7.37
asaltante	7.30	cable	6.91	arcas	7.36
politicos	6.36	bebé	6.82	banco	7.20
gato	3.68	armas	6.77	comercio	7.14
perro	3.64	cámara	6.63	transeúnte	7.12
		combustible	6.55	turista	6.90
		ganado	6.08	joyería	6.85
		comida	6.03	cajero	6.54

Table 20. (continued)

Robar algo a alguien/en o de un lugar / robar un lugar					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
		gallina	6.02	<i>casa</i>	6.27
		ropa	5.90	<i>tienda</i>	5.39
		mochila	5.87		
		animales	5.63		

Note: Collocates in italics are realized as direct objects, contrasting with the remaining ones that are typically realized as indirect objects or adjuncts in the form of PP introduced by *a*, *de* and/or *en*.

Table 21. Semantic categories of *robar*

Robar algo a alguien/en o de un lugar / robar un lugar					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	68	VEHICLE	12	HUMAN	51
ANIMAL	32	ASSETS	11	PLACE	49
		ORNAMENT	10		
		CASE	9		
		ELECTRONICS	9		
		HUMAN	9		
		WEAPON	9		
		FUEL	9		
		FOOD	8		
		ANIMAL	8		
		GARMENT	8		

The match between *rob* and *atracar* (Tables 22, 23, 24 and 25) is also partial because of their different syntactic complementation patterns resulting from their underlying semantics. *Rob* participates in one syntactic environment where the victim or source and the goods can both be stated (*rob someone/a place of something*). In contrast, *atracar* only refers to the source or victim of the robbery event (*atracar a alguien/un lugar*). With regards to the meaning definition of *rob*, the category ASSETS is found to be the most common type of goods robbed of.

Table 22. Lexical collocates of *rob*

<i>Rob someone of something / rob a place</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
thief	6.48	handbag	6.43	bank	8.60
gang	6.14	cash	5.42	store	6.85
bandit	4.81	wallet	4.33	grave	6.27
robber	4.33	cellphone	3.94	passenger	5.55
gunman	4.14	bike	3.94	tourist	5.49
suspect	4.07	purse	3.85	shopkeeper	5.23
		passport	2.78	supermarket	5.09
				citizen	4.51

Table 23. Semantic categories of *rob*

<i>Rob someone of something / rob a place</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	100	ASSETS	26	PLACE	56
		CASE	23	HUMAN	44
		VEHICLE	19		
		ELECTRONICS	19		
		DOCUMENT	13		

Table 24. Lexical collocates of *atracar*

<i>Atracar a alguien/un lugar</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
encapuchado	7.52			joyería	8.69
delincuente	5.84			sucursal	8.08
ladrón	5.64			gasolinera	7.78
asaltante	5.42			farmacia	7.08
individuo	4.69			banco	6.94
sujeto	4.42			supermercado	6.49
				casino	5.73
				taxista	5.51
				transeúnte	2.82
				pasajero	1.42

Table 25. Semantic categories of *atracar*

<i>Atracar a alguien/un lugar</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	100			PLACE	69
				HUMAN	31

As previously mentioned, our corpus study uncovers a number of discrepancies that require further evaluation. For instance, drawing on dictionaries and/or Google searches becomes a necessary step to establish cross-linguistic equivalence when examples are absent in the corpora. That is the case of *clean out* and *limpiar*, *descantillar* ‘embezzle’, and *pillar* ‘pillage’. However, in the case of *pisar* ‘steal’, we are unable to retrieve any information as to its theft meaning. When dictionaries do not provide sufficient information, we perform a Google search to evaluate the total equivalence of lexical collocates or semantic categories across languages. That is the case of *pilfer a document*,¹ *plunder petrol*,² *ratear tabaco*³ ‘pilfer tobacco’, or *birlar petróleo*⁴ ‘filch petrol’. In those instances where no equivalent or near-equivalent item is found for a given verb (e.g. *thieve*, *shoplift*, *swipe*, *loot*, *hurtar* ‘steal’, *sisar* ‘steal’), we compare that item to the most immediate superordinate member of the other language.

After careful examination of all verbs of theft, we find that the average cross-linguistic correspondence in the lexical domain of verbs of stealing roughly amounts to 70%, which is within the nominal scales of partial and total equivalence (Table 26).

Theft and robbery thus appear to be conceptualized more or less similarly in English and Spanish. Overall, the few differences between English and Spanish verbs of stealing are mainly related to different constructional patterns or slightly different lexical collocates and/or semantic categories. In this sense, our results differ from those of Enghels and Whylin (2015) and Dux (2011, 2018), who address the constructional patterns of verbs of stealing in French and Spanish and in English and German, respectively. To the best of our knowledge, no study

1. <https://www.easthempfield.org/2268/Identity-Theftwhat-can-you-do-to> [last accessed 28 December 2018]
2. <https://www.dailymail.co.uk/news/article-2083345/The-menace-gangs-using-drills-plunder-petrol.html> [last accessed 28 December 2018]
3. <https://www.forocoches.com/foro/showthread.php?t=5396291> [last accessed 28 December 2018]
4. <http://www.gees.org/articulos/testimonio-acerca-del-informe-volcker-del-petroleo-por-alimentos-de-la-onu> [last accessed 28 December 2018]

Table 26. Cross-linguistic correspondence matching rates of English and Spanish verbs of theft

English verb	Spanish verb	Matching equivalence
steal	robar	Partial: 75%
rob	atracar	Partial: 75%
hold up/stick up	asaltar	Total: 100%
thieve	robar	Partial: 50%
embezzle	desfalcar/descantillar	Partial: 75%
steal	hurtar	Partial: 75%
steal	sisar	Partial: 75%
purloin	sustraer	Partial: 75%
shoplift	robar	Partial: 50%
pilfer	ratear	Partial: 75%
pinch	mangar	Partial: 75%
steal	pisar	N/A
filch	birlar	Partial: 75%
lift	soplar	Partial: 75%
nick	afanar	Partial: 75%
swipe	robar	Partial: 50%
rustle	robar	Partial: 50%
clean out	limpiar	Partial: 75%
steal	desvalijar	Partial: 50%
make off with	arramblar	Total: 100%
plunder	saquear	Total: 100%
pillage	pillar	Partial: 50%
despoil	saquear	Partial: 75%
loot	saquear	Total: 100%

has provided an extensive list of the selection preferences in the form of lexical collocates and semantic categories for verbs of stealing in English and Spanish, let alone a corpus-based, contrastive study of these verbs.

Any outcome derived from a corpus-based, cross-linguistic study such as this has potential implications for cross-linguistic generalizations in ontology building, insofar as lexico-semantic issues are concerned. These implications relate to issues such as the following: (i) zero lexical equivalents; (ii) partial equivalence across languages; (iii) the representation of conceptual meaning.

For ontology building, the first case scenario demands the creation of a concept whose features and specifications should be universally applicable, despite the lack of lexicalization in a given number of languages (Espinoza *et al.*, 2009). This would apply to lexical items such as *shoplift* or *rustle* that have no direct translation equivalent in Spanish. In such cases, it would then be necessary to

create the concepts SHOPLIFT and RUSTLE and link them to their designations when they are lexicalized. In an ontology based on deep semantics, concepts must not only relate to other concepts in different types of semantic relations, but also specify the common conceptual-cognitive patterns and events associated with the participants that take part in them and their semantic nature (Periñán and Arcas, 2007; Velardi *et al.*, 1991). This indeed involves defining the most typical semantic categories attributed to each participant in the form of conceptual selection preferences.

In the second-case scenario where selection preferences might not coincide, ontology builders would have to check whether partial equivalence is due to the lack of exact lexicosemantic equivalence (i.e. sharing of the same semantic categories). If so, it would then be necessary to verify whether finding more abstract semantic categories could result in a complete match. Only then could both lexical items be linked to a given concept. For example, pinching *money* or *cash* in the sense of stealing might be ungrammatical because it has not been evidenced in the corpora, dictionaries or on the Web. On the other hand, *mangar dinero* or *pasta* ‘pinch money’ is perfectly acceptable according to corpus evidence. From this, we could conclude that the ASSETS category for goods is not present in *pinch* but rather in its Spanish equivalent *mangar*. However, ASSETS can be subsumed within the more general ARTEFACT category. The next step is to ensure that ARTEFACT for goods coincides in both verbs, which is indeed the case. More specifically, CASE, COSMETICS, GARMENT, and VEHICLE are ARTEFACT for goods in *pinch* (Tables 27 and 28), and GARMENT, ASSETS, ELECTRONICS, DOCUMENT, and CONTAINER are ARTEFACT for goods in *mangar* (Tables 29 and 30).

Table 27. Lexical collocates of *pinch*

<i>Pinch something from someone/a place</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
bugger	2.41	wallet	5.17	Tories	2.37
someone	1.86	purse	4.31	brother	1.87
		razor	2.11	people	1.77
		apple	2.07	house	1.73
		handbag	1.99	coffer	1.38
		t-shirt	1.92	car	0.38
		bicycle	1.89		
		cake	1.84		
		cigarette	1.73		
		car	1.12		

Table 28. Semantic categories of *pinch*

<i>Pinch something from someone/a place</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	100	CASE	29	HUMAN	63
		COSMETICS	16	PLACE	37
		FOOD	15		
		GARMENT	15		
		DRUGS	13		
		VEHICLE	12		

Table 29. Lexical collocates of *mangar*

<i>Mangar algo a alguien/de un lugar</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean	Lexical collocates	LogDice score mean
chachos	N/A	manuales	N/A	tienda	N/A
izquierda	N/A	ordenadores	N/A	gente	N/A
		discos	N/A	hotel	N/A
		pasta	N/A		
		dinero	N/A		
		higos	N/A		
		jamones	N/A		
		botellas	N/A		
		antena	N/A		
		calzoncillos	N/A		
		ropa	N/A		
		guantes	N/A		

Table 30. Semantic categories of *mangar*

<i>Mangar algo a alguien/de un lugar</i>					
PERPETRATOR		GOODS		VICTIM/SOURCE	
Semantic categories	%	Semantic categories	%	Semantic categories	%
HUMAN	N/A	GARMENT	N/A	PLACE	N/A
		ASSETS	N/A	HUMAN	N/A
		ELECTRONICS	N/A		
		FOOD	N/A		
		DOCUMENT	N/A		
		CONTAINER	N/A		

The absence of DRUGS in *mangar* can be explained by the lack of corpus examples, since we can find one instance of *mangar cigarros*⁵ ‘pinch cigarettes’ on the Web. As a result, both lexical items can be claimed to refer to the same generic concept, that of STEAL. In this sense, minor differences in the lexical selection preferences of particular verbs should be handled in the lexicon component of a given ontology.

In the third case scenario, the creation of new concepts should be handled by identifying those semantic discrepancies in the meaning components obtained from dictionary and encyclopedic knowledge in the LGM hierarchy. In the case of verbs of stealing, attention should be paid, not only to their selection preferences, but also to their differentiating meaning characteristics. In the LGM hierarchy, verbs such as *plunder* and *saquear* and their hyponyms are considerably more specific in their semantics than *steal* and other subordinate verbs. This would demand the creation of a separate concept PLUNDER because of the encyclopedic knowledge associated with the context in which plundering occurs (i.e. in violent activities or events, in which great damage is caused and valuable items are typically taken).

4.3 Word sketches: Issues and limitations

The use of word sketches comes with a few caveats that any study of the selection preferences of other verbal domains should take into account. One of these caveats has to do with the productive character of the lexicon by means of polysemy and the lexico-semantic phenomena of metaphoric and metonymic extension associated with it (Asher, 2011).

The verbs at the top in the lexical hierarchy, which display a wider semantic scope, are typically more prone to metaphorical and metonymic extension. For this reason, it is no coincidence that idiom-like, metaphorical and other secondary meanings widely populate word sketches with their collocational patterns in more generic verbs such as *steal* or *rob* (e.g. *steal a march*, *steal a glance*, *steal the show*, *steal somebody’s thunder*, *rob Peter to pay Paul*...). Other more specific stealing verbs (e.g. *clean out*) also raise other issues in word sketches. For instance, word sketches of *clean out* show that its stealing meaning is hidden under a much larger number of lexical collocates corresponding to the meaning of making something neat. Other more generic verbs in the POSSESSION domain such as *have* or *give* would present further challenges because of their higher polysemous nature.

5. http://blogs.montevideo.com.uy/blognoticia_10877_1.html [last accessed 28 December 2018]

Another problem encountered with the use of word sketches is the presence of corpus noise. For instance, sometimes the syntactic parser does not recognize lemmas, resulting in the multiple presence of inflected morphological forms in the lists of lexical collocates, each with a different logDice salience score. For instance, the lemma *delincuente* ‘criminal’ appears twice, in the singular with a 9.02 logDice salience score and in the plural with 5.05 logDice salience score, both in the subject list of the collocates of *robar* ‘steal’ in the esTenTen11 corpus. In these cases, we pick the collocate with the highest score for the calculation of the mean. At other times, we can find wrongly assigned lexical collocates due to an erroneous syntactic parsing – *horse stealing* in the subject list of collocates rather than in the object list – or words that do not exist, for instance, the collocate *n23* in the object list of the collocates of *embezzle*. This is particularly frequent at the bottom of the lists.

All this highlights the need for manual intervention in order to discard corpus noise and lexical collocates that correspond to secondary, metaphorical or idiom-like senses. The only way to do this is to click on dubious lexical collocates and analyze actual corpus instances to ensure that lexical collocates have been correctly extracted.

5. Conclusion

We have presented a contrastive study of English and Spanish verbs of stealing that is based on the Lexical Grammar Model enriched with certain premises of Frame Semantics. Our research specifies the semantic categories for the arguments of stealing verbs and analyzes their potential cross-linguistic correspondence at both semantic and syntactic levels. To the best of our knowledge, no study has used the Word Sketch tool to carry out cross-linguistic studies of selection preferences, or has focused on verbs of stealing from a semantic, corpus-driven perspective. Only a few recent studies have analyzed verbs of stealing in English, Spanish, French or German, though from a purely constructional point of view.

The Word Sketch tool, despite certain issues and limitations, remains one of the most useful and robust semi-automatic corpus search tools for the extraction of lexical collocates. It would thus be possible to extrapolate and apply the methodology presented in our study to the analysis of the selection preferences of verbs belonging to other semantic domains in one or several languages.

The results of our study provide a comprehensive account of the selection preferences of verbs of stealing in English and Spanish and their degree of inter-linguistic correspondence. We have presented the general semantic categories of participants in the conceptual scenario of theft and robbery and discussed the

differences and similarities between languages. This could pave the way for future research on the conceptual scenarios for other semantic domains, as well as the semantic types of the participants involved.

The corpus-based, cross-linguistic procedure applied in our study could also contribute to the development of more fine-grained ontologies based on deep semantics with language-independent concepts that reflect not only semantically related concepts, but also their cognitive-conceptual space through the semantic categories of their participants. This would greatly benefit the performance of NLP tasks such as word sense disambiguation, machine translation, semantic tagging and semantic parsing.

Acknowledgements

This research was carried out as part of the research project *Herramientas terminológicas orientadas hacia la traducción de textos medioambientales* (FFI2017-89127-P), funded by the Spanish Ministry of Economy and Competitiveness.

References

- Asher, N. 2011. *Lexical Meaning in Context: a Web of Words*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511793936>
- Baisa, V., Moze, S. and Renau, I. 2016. Linking Verb Patterns Dictionaries of English and Spanish. *Proceedings of the Seventeenth European Association for Lexicography International Congress EURALEX '16*. Tbilisi, Georgia, 6–10 September 2016. Ivane Javakhishvili Tbilisi State University. 410–417.
- Barsalou, L. 1992. Frames, Concepts and Conceptual Fields. In *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, E. Kittay and A. Lehrer (eds), 21–74. Hillsdale: Lawrence Erlbaum Associates.
- Berman, R. 1982. On the Nature of 'Oblique' Objects in Bitransitive Constructions. *Lingua* 56(2): 101–125. [https://doi.org/10.1016/0024-3841\(82\)90026-2](https://doi.org/10.1016/0024-3841(82)90026-2)
- Boas, H. 2005. Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. *International Journal of Lexicography* 18(4): 445–478. <https://doi.org/10.1093/ijl/ecio43>
- Boas, H. 2008. Towards a Frame-Constructional Approach to Verb Classification. *Revista Canaria de Estudios Ingleses* 57: 17–48.
- Boas, H. 2013. Frame Semantics and Translation. In *Cognitive Linguistics and Translation*, A. Rojo and I. Ibarretxe-Antunano (eds), 125–158. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110302943.125>
- 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 28 December 2018].

- Butler, C. 2003. *Structure and Function: a Guide to Three Major Structural-Functional Theories. Part I: Approaches to the Clause*. Amsterdam: John Benjamins.
- Cambridge Learner's Dictionary. 2018. Cambridge: Cambridge University Press. Available at <https://dictionary.cambridge.org/es/> [last accessed 28 December 2018].
- Collins Free Online Dictionary. 2018. Glasgow: HarperCollins Publishers. Available at <https://www.collinsdictionary.com/es/> [last accessed 28 December 2018].
- Diccionario de la Real Academia Española (DRAE). 2018. Madrid: Real Academia Española. Available at <http://www.rae.es/> [last accessed 28 December 2018].
- Diccionario Salamanca de la Lengua Española. 2018. Madrid: Santillana Educación. Available at <http://fenix.cnice.mec.es/diccionario/> [last accessed 28 December 2018].
- Dik, S. 1978. *Functional Grammar*. Dordrecht: Foris Publications.
- Dux, R. 2011. A Frame-Semantic Analysis of Five English Verbs evoking the Theft Frame. Master's Dissertation, University of Texas. Available at <https://repositories.lib.utexas.edu/handle/2152/ETD-UT-2011-05-3114> [last accessed 28 December 2018].
- Dux, R. 2018. Frames, Verbs, and Constructions: German Constructions with Verbs of Stealing. In *Approaching German Syntax from a Constructionist Perspective*, A. Ziem and H. Boas (eds), 367–405. Berlin: Mouton de Gruyter.
- Engels, R. and Wylin, K. 2015. Expressing the Source of Dispossession Acts in French and Spanish. *Languages in Contrast* 15(1): 102–124. <https://doi.org/10.1075/lic.15.1.06eng>
- Espinoza, M., Montiel-Posoda, E. and Gómez-Peréz, A. 2009. Ontology Localization. *Proceedings of the Fifth International Conference on Knowledge Capture K-CAP '09*. Redondo Beach, California, USA, 1–4 September 2009. ACM. 33–40. <https://doi.org/10.1145/1597735.1597742>
- Faber, P. and Mairal, R. 1998a. Prototipos semánticos en el lexicon Lexemático Funcional. In *Estudios de tipología lingüística*, J. D. D. Luque-Durán and A. Pamies-Bertrán (eds), 15–36. Granada: Método.
- Faber, P. and Mairal, R. 1998b. Towards a Typology of Predicate Schemata in a Functional-Lexematic Model. In *Towards a Functional Lexicology*, G. Wotjak (ed), 11–37. Frankfurt: Peter Lang.
- Faber, P. and Mairal, R. 1998c. Towards a Semantic Syntax. *Revista Canaria de Estudios Ingleses* 36: 37–64.
- Faber, P. and Mairal, R. 1999. *Constructing a Lexicon of English Verbs*. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110800623>
- Faber, P. and Mairal, R. 2017. A Conceptually-Oriented Approach to Semantic Composition in RRG. In *The Cambridge Handbook of Role and Reference Grammar*, R. D. Van Valin (ed). Cambridge: Cambridge University Press.
- Fillmore, C. 1982. Frame Semantics. In *Linguistics in the Morning Calm*, Linguistic Society of Korea (ed), 111–137. Seoul: Hanshin Publishing.
- Fillmore, C. 1985. Frames and the Semantics of Understanding. *Quaderni di Semantica* 6: 222–254.
- Fillmore, C., Baker, C. and Lowe, J. 1998. The Berkeley FrameNet Project. *Proceedings of the Seventeenth International Conference on Computational Linguistics COLING '98*. Montreal, Quebec, Canada, 10–14 August 1998. Université de Montréal. 86–90. Available at <https://framenet2.icsi.berkeley.edu/> [last accessed 28 December 2018].
- Fillmore, C. and Baker, C. 2010. A Frames Approach to Semantic Analysis. In *The Oxford Handbook of Linguistic Analysis*, B. Heine and H. Narrog (eds), 313–340. Oxford: Oxford University Press.


- Gilardi, L. and Baker, C. 2018. Learning to Align across Languages: Toward Multilingual FrameNet. *Proceedings of the International FrameNet Workshop 2018: Multilingual Framenets and Constructicons*. Miyazaki, Japan, 12 May 2018. 13–22.
- Goldberg, A. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.
- Goldberg, A. 2010. Verbs, Constructions and Semantic Frames. In *Syntax, Lexical Semantics and Event Structure*, M. Rappaport-Hovav, E. Doron and I. Sichel (eds), 39–58. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199544325.003.0003>
- Gruzitis, N. and Dannéls, D. 2017. A Multilingual FrameNet-based Grammar and Lexicon for Controlled Natural Language. *Languages Resources and Evaluation* 51(1): 37–66. <https://doi.org/10.1007/s10579-015-9321-8>
- Hanks, P. 2004. Corpus Pattern Analysis. *Proceedings of the Eleventh European Association for Lexicography International Conference EURALEX '04*. Lorient, France, 6–10 July 2004. Université de Bretagne-Sud. 87–97.
- Hanks, P. 2012. How People Use Words to Make Meanings: Semantic Types Meet Valencies. In *Input, Process and Product: Developments in Teaching and Language Corpora*, A. Boulton and J. Thomas (eds), 52–67. Brno: Masaryk University Press.
- Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge: MIT Press. <https://doi.org/10.7551/mitpress/9780262018579.001.0001>
- Hanks, P. and Jezek, E. 2010. What Lexical Sets Tell Us about Conceptual Categories. *Lexis 4: Corpus Linguistics and the Lexicon* 7–22.
- Harley, H. 2003. Possession and the Double Object Construction. In *Linguistic Variation Yearbook 2*, P. Pika and J. Rooryck (eds), 31–70. Amsterdam: John Benjamins.
- Heine, B. 1997. *Possession: Cognitive Sources, Forces and Grammaticalization*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511581908>
- Jakubíček, M., Kilgarrieff, A., Kovář, V., Rychlý, P. and Suchomel, V. 2013. The TenTen Corpus Family. In *Corpus Linguistics 2013: Abstract Book*, A. Hardie and R. Love (eds), 125–127. Lancaster: UCREL.
- Kilgarrieff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. 2014. The Sketch Engine: Ten Years on. *Lexicography* 1: 7–36. Available at https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2014.pdf [last accessed 28 December 2018].
- Kilgarrieff, A. and Renau, I. 2013. esTenTen, a Vast Web Corpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences* 95: 12–19. <https://doi.org/10.1016/j.sbspro.2013.10.617>
- Kilgarrieff, A., Vojtěch, K., Krek, S., Srdanović, I. and Tiberius, C. 2010. A Quantitative Evaluation of Word Sketches. *Proceedings of the Fourteenth European Association for Lexicography International Congress EURALEX '10*. Leeuwarden, Netherlands, 6–10 July 2010. Fryske Akademy. 372–379.
- Langacker, R. 2007. Cognitive Grammar. In *The Oxford Handbook of Cognitive Linguistics*, D. Geeraerts and H. Cuyckens (eds), 421–462. Oxford: Oxford University Press.
- Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.
- Longman Dictionary of Contemporary English. 2018. London: Pearson Longman. Available at <https://www.ldoceonline.com/> [last accessed 28 December 2018].
- Marsh, I., Melville, G., Morgan, K., Norris, G. and Walkington, Z. 2006. *Theories of Crime*. London: Routledge.

- McCarthy, D., Kilgarriř, A., Jakubiček, M. and Reddy, S. 2015. Semantic Word Sketches. In *Corpus Linguistics 2015: Abstract Book*, F. Formato and A. Hardie (eds), 231–233. Lancaster: UCREL.
- Miller, G. and Fellbaum, C. 2007. WordNet Then and Now. *Language Resources and Evaluation* 41(2): 209–214. Available at <https://wordnet.princeton.edu/> [last accessed 28 December 2018].
- Oxford English Dictionary. 2018. Oxford: Oxford University Press. Available at <https://en.oxforddictionaries.com/> [last accessed 28 December 2018].
- Periñán, C. and Arcas, F. 2007. Deep Semantics in an NLP Knowledge Base. Paper presented at the Twelfth Conference of the Spanish Association for Artificial Intelligence CAEPIA '07, Spain, 12–16 November.
- Ruppenhofer, J., Boas, H. and Baker, C. 2017. FrameNet. In *The Routledge Handbook of Lexicography*, P. Fuertes-Olivera (ed), 383–398. London: Routledge.
- Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. and Scheffczyk, J. 2010. *FrameNet II: Extended Theory and Practice*. Berkeley: International Computer Science Institute.
- Rychlý, P. 2008. A Lexicographer-Friendly Association Score. *Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Languages Processing RASLAN '08*. Karlova Studánka, Czech Republic, 5–7 December 2008. Masaryk University. 6–9.
- Subirats, C. 2009. Spanish FrameNet: A Frame Semantics Analysis of the Spanish Lexicon. In *Multilingual FrameNets in Computational Lexicography. Methods and Applications*, H. Boas (ed), 135–162. Berlin: Mouton de Gruyter. Available at <http://spanishfn.org/> [last accessed 28 December 2018].
- 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.
- Trampus, M. and Novak, B. 2012. The Internals of an Aggregated Web News Feed. Paper presented at the Fifteenth Multiconference on Information Society, Slovenia, October 2012.
- Velardi, P., Pazienza, M. and Fasolo, M. 1991. How to Encode Semantic Knowledge: a Method for Meaning Representation and Computer-aided Acquisition. *Computational Linguistics* 17(2): 153–170.
- WordReference Dictionary. 2018. Available at <http://www.wordreference.com/> [last accessed 28 December 2018].

Address for correspondence

Nicolás José Fernández-Martínez
 Department of English Language and Linguistics
 University of Granada
 Campus de la Cartuja, Universidad de Granada, Prof. Clavera
 Granada, 18011
 Spain

njfm0001@red.ujiaen.es

 <https://orcid.org/0000-0002-0650-7625>

Co-author information

Pamela Faber
Department of Translation and Interpreting
University of Granada
pfaber@ugr.es

Publication history

Date received: 9 January 2019
Date accepted: 17 May 2019
Published online: 5 June 2019