RAEL: Revista Electrónica de Lingüística Aplicada

Vol./Núm.:	21/1
Enero-diciembre	2022
Páginas:	115-128
Artículo recibido:	30/06/2022
Artículo aceptado:	06/11/2022
Artículo publicado	31/01/2023

Url: https://rael.aesla.org.es/index.php/RAEL/article/view/495
DOI: https //doi.org/10.58859/rael.v21i1.495

Extracting and Formalizing Criminal-law Terminology Using DEXTER and FunGramKB

Extracción y formalización de terminología del derecho penal con la herramienta DEXTER y FunGramKB

ÁNGELA ALAMEDA HERNÁNDEZ
PEDRO UREÑA GÓMEZ-MORENO
UNIVERSIDAD DE GRANADA

In this article we discuss the processes of terminology extraction and specialized knowledge conceptualization taking the field of criminal law as an illustrative case in point. Regarding the first process, we provide an overview of DEXTER, a tool that has proven effective in detecting specialized units from a collection of texts with minimal effort on the part of the user. Regarding the conceptualization phase, this article illustrates the conceptual definition of some representative units belonging to the field of *criminal offenders*. The purpose of conceptual modeling along these lines is the inclusion of specialized units of thought into a knowledge base known as FunGramKB. The aim of this paper is to show that the combination of an extractor such as DEXTER with the semantic expressiveness of FunGramKB constitutes a solid basis for the improvement of specialized technological tools aimed at jurists, legal specialists, or other practitioners interested in legal terminology.

Keywords: *terminological extraction; DEXTER; FunGramKB; criminal law*

Este trabajo ilustra los procesos de extracción de terminología y conceptualización del conocimiento especializado tomando como ejemplo unidades del campo del derecho penal. Con respecto al primer proceso, brindamos una descripción general de DEXTER, una herramienta que ha demostrado ser efectiva para detectar unidades especializadas de una colección de textos con un mínimo esfuerzo por parte del usuario. En cuanto a la fase de conceptualización, este artículo ilustra la definición conceptual de algunas unidades representativas pertenecientes al campo de los *infractores penales*. El propósito del modelado conceptual en este sentido es la inclusión de conceptos especializados en una base de conocimiento conocida como FunGramKB. El objetivo de este trabajo es mostrar que la combinación de un extractor como DEXTER con la expresividad semántica de FunGramKB constituye una base sólida para la mejora de herramientas tecnológicas especializadas dirigidas a juristas u otros profesionales interesados en la terminología jurídica.

Palabras clave: *extracción terminológica; DEXTER; FunGramKB; derecho penal*

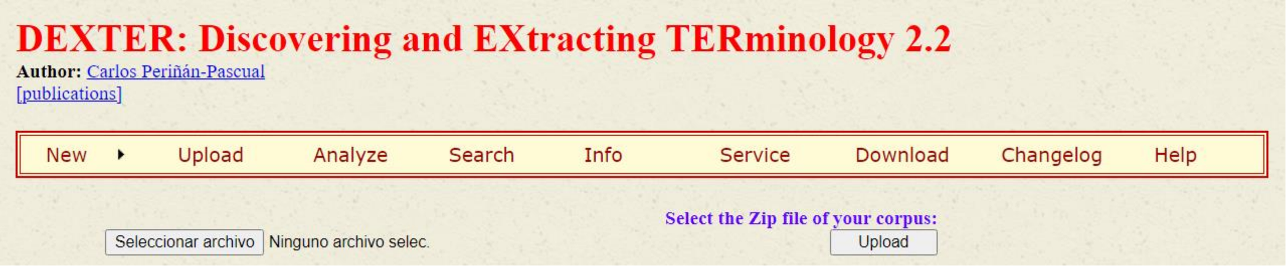
1. Introduction

The 21st century is witnessing an exponential growth in the number of documents available in almost any area of expertise. While this circumstance is making a decisive contribution to the dissemination of scientific and technological content worldwide, it also poses a challenge for professionals who may require access to textual information in a fast and straightforward way. New technologies can contribute to deal with such a vast number of digital documents and, indeed, this has been the focus of much academic research in the last decades (Ikonomakis, Kotsiantis & Tampakas, 2005; Bárcena, Read & Arús Hita, 2014; Chu, 2014; Rajni, Ruchika & Jain, 2015; Lausch, Schmidt & Tischendorf, 2015). Likewise, advances in text mining and document management have led to the emergence of several tools for accomplishing a variety of language-related tasks, including topic identification, document classification, text summarization or human-machine interaction. All these resources have definitely helped to alleviate the textual deluge; however, there is still a need to solve the problem of how specialized documentation should be computationally treated so that it can be more accessible both to experts and non-experts. An important step in resolving this problem lies in the work of *terminology*—traditionally defined as an academic discipline which involves the study, compilation and organization of specialized terms (henceforth, terminology; Cabré, 1999)—plays a fundamental role in this scenario.

If mastery of expert knowledge is essential for professionals in a given field of expertise, it naturally follows that terminology should also be an important element of specialized linguistic resources, such as online glossaries, automated translators, or domain-specific ontologies. The question lies in how to identify terminology in a way that is relevant for any area of human knowledge (Geology, Biology, Engineering, Physics, etc.) and that takes the minimum effort and time to achieve. In this regard, *Terminology extraction* is a consolidated area both within Terminology and Natural Language Processing (NLP), which is essential in providing the necessary algorithms for the automatic identification of terminology. Extraction can be therefore defined as the computational process by which specialized concepts are retrieved automatically from a collection of texts. This paper focuses on terminological extraction applied to the field of criminal law, with the goal of showing how to incorporate terms retrieved from this legal domain into a knowledge base called FunGramKB (Periñán-Pascual & Arcas-Túnez, 2010; Periñán-Pascual & Mairal-Usón, 2010). *Knowledge base* can be defined as a machine-readable repository which contains concepts representing any entity, event or quality in the world (for example, *Human, Law, Go, Tired*, etc.), as well as relevant lexico-grammatical information associated to the lexical units that represent those concepts (for example, FunGramKB establishes that the concept *Tired* is universal to all languages, because every concept we can imagine has an appropriate place in the ontology (Periñán-Pascual & Arcas-Túnez, 2010: 2668), but each language uses a different word to represent it: “tired” in English, “cansado” in Spanish, o “müde” in German). This knowledge base is intended for use in NLP and other applications where logical and linguistic reasoning is required, such as the creation of artificial intelligence systems. For this goal, we propose a methodology in four phases: a) text collection, b) automatic extraction, c) terminological filtering, and d) conceptualization. The proposed method can be used to improve already existing terminological banks, as well as to populate ontological repositories. Within the area of criminal law, we offer illustrative examples on how to conceptualize different types of offenders, that is, human agents who have committed a crime.

For the extraction phase mentioned above, we propose DEXTER (i.e., Discovering and EXtracting TERminology), a term-mining tool which offers a great potential for the detection, as well as for the filtering of lexical units. The first version of the extractor was called

"FunGramKB Term Extractor" and was implemented and launched on the web with the purpose of providing terminologists and practitioners with a more robust tool for terminology identification than was offered by other existing extractors (Felices-Lago & Ureña Gómez-Moreno, 2014). The main advantages of this early version of the extractor consisted in a great capacity for large corpora processing, as well as in offering several key tools for terminology management. In addition, the tool showed some additional advantages in terms of extraction capacity, as well as in assisting the terminologist in the decision-making phase of which units must be considered specialized. Just recently, however, the extractor has undergone a major redesign of its processing engine along with some other improvements in the interface and the user-interaction flow. Due to these essential changes in the tool, it has been renamed as DEXTER (Periñán-Pascual, 2018). DEXTER is available online to anyone interested in terminology or languages for specific purposes (http://www.fungramkb.com/), and on this principle the interface has been designed taking into account the different types of potential users. In this regard, the tool shows a minimalist interface, especially designed for students or scholars who are not familiar with terminology extraction, whilst it also allows to customize advanced parameters related to the statistical phase. To start with the extraction, the first step is to create a new corpus in the *New* section (Figure 1). On this menu the user will be asked to upload the texts in batches, and to provide a brief comment on the name or content of the texts. The next step consists in the analysis of the uploaded sample using the *Analyze* menu (Figure 1), where both the linguistic filters and the statistical measure are applied in order to retrieve the list of potential terminological units. For more information on the different types of tool utilities, the reader is referred to Felices-Lago and Ureña Gómez-Moreno (2014).



DEXTER: Discovering and EXtracting TERminology 2.2

Author: Carlos Periñán-Pascual

[publications]

Figure 1: ***Main panel of DEXTER, a tool for term extraction***

The remainder of this paper is structured as follows. Section 2 defines and summarizes the main aspects of knowledge formalization within the framework of FunGramKB. Section 3 delves into the process of terminology extraction and conceptual modeling. Finally, Section 4 illustrates the process of conceptual definition of entities by offering a sample of several terms belonging to the field of *criminal offenders*.

2. Knowledge formalization in the domain of Law

The conceptualization of legal knowledge poses many difficulties to *knowledge engineers*, i.e., scholars who study the semantics of terms with the aim of translating them into concepts using a formalization language. The following claim by Liebwal (2007: 140) states the general problem in a very explicit way: "[…] the cross-linking of different domain models and the interconnection of the concept spaces of world knowledge (the world model) and legal knowledge (the domain models) are still substantial problems." Moreover, there are some additional difficulties—some of which are also mentioned in Liebwald (2007)—in the conceptualization of both basic and highly specialized legal concepts. For example, many of the terminology which is recurrently used in different types of legal documents, such as sentences, warrants or court orders, show a great deal of semantic complexity, which makes these lexical units difficult to synthesize into a machine-readable language. Furthermore, despite the fact that there are concepts which are shared among different legal systems and can be considered universal, the conceptualization of other processes, events, participants, etc. are dependent on (or are constrained by) the specific legal system of a country, or even its language and culture.

The field of Law has been a particularly interesting area for the development of terminological studies as well as domain ontologies (Valente, 2005; Breuker, Casanovas, Klein & Francesconi, 2009; Casanovas, Sartor, Biasiotti & Fernández-Barrera, 2011). Ontologies are machine-readable repositories which contain concepts representing all the knowledge (participants, events, features, relations, etc.) in a particular domain. Consequently, these ontologies are key components, for example, in artificial intelligence applications in which legal reasoning is involved. What is interesting is that ontologies require a terminological basis to be implemented successfully. We must emphasize that despite the numerous advances in this field of study, some of the proposed ontologies do not have a fully computational implementation and/or lack a model based on deep semantics and the linking between linguistic and conceptual modules as in FunGramKB. In this paper, we focus our attention on the core ontology in FunGramKB, a lexico-conceptual knowledge base for NLP, and the Satellite (terminological) ontologies linked to it.

FunGramKB is multipurpose, because it can be used in a variety of computational tasks, such as information retrieval, machine translation or artificial reasoning, and also because it has been designed to work with any human language (Periñán-Pascual & Arcas-Tunez, 2010). Whilst other knowledge bases are grounded on surface semantics and lexical relations, FunGramKB relies on deep semantics, i.e., it stores fully-fledged representations of conceptual units. As for its architecture, FunGramKB is structured into three modules which represent three different levels of knowledge: a) lexical; b) grammatical; and c) conceptual. These levels are independent, although interrelated. The ontology is the central component of the conceptual module and it is conceived as a taxonomy of concepts that represent general knowledge. Concepts, and not words, are the building blocks for the formal representation of other concepts, so that these become language-independent semantic representations. This is possible thanks to the use of a formal representation language named COREL, which is conceived as a conceptual representation notation system (Periñán-Pascual & Mairal-Usón, 2010). All these features make FunGramKB particularly suitable for developing applications to aid experts in the legal field, who need to face many time-consuming tasks involved in dealing with legal documents.

The need to expand FunGramKB's *Core Ontology*, which contains only common-sense concepts, and integrate the so-called *Satellite Ontologies* of specialized knowledge was originally proposed by Periñán-Pascual (2010) and Felices-Lago and Ureña Gómez-Moreno (2012) and was later substantiated in the construction of two Satellite Ontologies: a) one for the specific domain of organized crime and terrorism (Carrión-Delgado & Felices-Lago, 2014; Periñán-Pascual & Arcas-Túnez, 2014; Felices-Lago, 2015; Alameda-Hernández & Felices-Lago, 2016; Ureña Gómez-Moreno, 2016; Alameda-Hernández and Felices-Lago, 2017); and b) a second one, incipiently developed, in the aviation domain (Felices-Lago & Alameda-Hernández, 2017). Following on from the successful implementation of those Satellite Ontologies, the present research has led us to the development of a specialized ontology to cover the full and broader field of criminal law. It is also relevant to highlight a further reason why this specialized ontology is going to be linked to the well-developed core ontology of FunGramKB—instead of being created as a separated or autonomous system—as San Martín and Faber (2014) stated, lacking a link to general knowledge could lead to incompleteness, incoherence or ambiguity in the domain-specific conceptual relations.

The Core Ontology, as a hierarchical repository of concepts, is organized into three broad categories or subontologies, namely ENTITY, EVENT and QUALITY, which permit the internal organization of lexical units: nouns, verbs, and adjectives, respectively. In this paper, we have selected concepts that belong to the category ENTITY, and more precisely, those with the "+human" semantic feature, that is, those human entities involved in criminal law proceedings. The reason is that even if criminal law centrally refers to the body of laws that define and apply to criminal offenses, the basic question eventually boils down to "who" is responsible for what crime (Samaha, 2011: 7). In addition, it has been noticed that in the recent years, a lot of effort has been put to populate and expand the Satellite Ontology of FunGramKB, but this work has mainly been directed to those entities representing crimes, but little attention has been given to the agents of those crimes.

3. A METHODOLOGY FOR TERMINOLOGY EXTRACTION AND CONCEPTUAL MODELING

This section presents the phases in the process of terminology extraction and the ontological definition of the selected terms. These phases are divided into: a) text collection; b) automatic term extraction; c) manual filtering; and d) conceptualization.

The first phase of the extraction process consists in compiling a collection of texts, which will be used as an input dataset to which DEXTER's statistical engine will be applied later. For this paper, a collection of texts containing approximately 1.7 million words was used as the evaluation corpus. In order to come up with a relevant set of texts we used a selection of keywords in the field: *criminal liability, mens rea, actus reus, concurrence* and *harm*, that a pilot study had proved to be particularly productive for the retrieval of relevant documents dealing specifically with criminal law. In the second phase, the collection of texts was uploaded and analyzed by DEXTER in order to retrieve a first set of candidate units. We define *candidate unit* as the lexical units that the machine extract from the dataset but have not yet been validated by a terminologist or a domain expert. For example, words such as *aggravate, discrete,* or *affirmation*. The candidates which are eventually admitted as fully terminological units will be called *winning candidates*. The results of our sample show a heterogeneous group of units, as well as other units which are semantically related to specific aspects of the criminal process. Such is the case of the verbs related to the commission of crimes (e.g., *impute, misappropriate*), the criminal acts themselves (e.g., *misdemeanor, larceny*), or the criminal agents (e.g. *aider, abettor*). The third stage, therefore, is to detect basic concepts that can hold other related concepts in a hyponymic relationship. That is the case of the selection for this paper, in which the concept "criminal" subsumes other more specific units such as "felon", "misappropriator", "impersonator" or "infringer."

Once this first terminological work is completed, the method involves a second step that consists in the conceptualization modeling. In order to integrate the selected terms into FunGramKB, Periñán-Pascual and Mairal-Usón (2011) proposed the COHERENT methodology, which consists of four phases: a) conceptualization; b) hierarchization; c) remodeling; and d) refinement. For the purpose of this paper, only the first two phases will be outlined since the remaining two phases fall outside the scope of this research. On the one hand, remodeling generally affects verbal concepts under the #EVENT subontology, but not entities. As for refinement, it is not pertinent at present since, it is only when the satellite ontology is sufficiently populated that this phase is required, as it will imply removing, merging, or demoting created concepts that have proved not productive.

Then, the conceptualization phase involves the process for the conversion of lexical units in a language into an inventory of general concepts shared by any of the languages used in the

FunGramKB lexical component. This requires a thorough analysis of lexicographical sources and, particularly for the purpose of building a Satellite Ontology, a study of specialized resources which will allow the researcher to eventually produce a final list of concepts in the expert domain. The hierarchization phase consists in allocating the concepts from the previous stage into any of the three subontologies that comprise the FunGramKB conceptual module: a) ENTITY, for nouns; b) EVENT, for verbs; and c) QUALITY, for adjectives. In the present paper, only entities involved in criminal events will be analyzed.

Concepts in FunGramKB are organized in an IS-A subsumption hierarchy (Periñan-Pascual & Arcas-Tunez, 2010), which means that each concept inherits the semantic properties of the concept above. Consequently, for each of the concepts that result from the previous phase, hypernymic and hyponymic relations with other concepts were identified. In addition, in this hierarchical organization, the ontology of this knowledge base identifies three conceptual levels which refer to different degrees of specificity (Periñan-Pascual, 2013: 90): a) metaconcepts; b) basic concepts; and c) terminal concepts. *Metaconcepts* are cognitive dimensions that conform the upper conceptual level and are preceded by the symbol "#" (e.g., #MOTION, #POSESSION). *Basic concepts* represent core knowledge and can be used as defining units for other concepts in the ontology. They are identified with the symbol "+" (e.g., +WINDOW_00, +WEAR_00, +CLEAN_00). Finally, *terminal concepts*, which are preceded by the symbol "$", are the final node in the hierarchy and they lack definitory potential (e.g., $EXCHANGE_00, $BARGAIN_00). Then, each concept was provided with semantic properties which are captured by Meaning Postulates (MP). These are a set of one or more logically connected predications (e1, e2, …), i.e., conceptual constructs that represent the generic semantic features of a concept. MPs are formalized using a representation language called COREL that has its own syntax and notation system. The most important feature of this language is that for the definition of a concept it uses other concepts and operators to establish the connections between them, which gives the ontology a considerable expressive power.

The list of candidate terms filtered by DEXTER must undergo manual inspection in order to identify the winning terms that could finally enrich FunGramKB with criminal law knowledge. Together with the initial work with both general and specialized dictionaries, a further analysis was carried out to identify which of these candidate terms were already included either in the Core Ontology or in the Satellite Ontology on organized crime and terrorism. Indeed, we found that several candidates retrieved by DEXTER were already included in FunGramKB, either as concepts or as lexical units associated to other concepts. Such was the case of "conspirator" and "fugitive", which are both present as terminal concepts $CONSPIRATOR_00 and $FUGITIVE_00, respectively in the Core Ontology, while they are also included in the Satellite Ontology of organized crime and terrorism as basic concepts (+CONSPIRATOR_00 and +FUGITIVE_00). These are examples of so-called "mirror concepts" (Carrión-Delgado, 2012), that is, concepts that, although being part of general knowledge in the Core Ontology, acquire a more precise meaning when used by experts and consequently need to the formalized in the Satellite Ontology with a more fine-grained conceptual definition. In addition, other candidates such as "accomplice" or "perpetrator" are already included in the ontology as lexical units linked to the basic concepts +CONSPIRATOR_00 and +CRIMINAL_00, respectively. Following from the analysis of this first selection of candidate terms it stems that all these units have been found to be already formalized as specialized knowledge in the ontology of FunGramKB, thus supporting the pertinence of the results offered by the DEXTER tool. Additionally, since all the conceptual modules of FunGramKB are linked, these candidates will not require any further conceptualization in a satellite ontology on criminal law, hence maximizing the reuse of conceptual information from the core as well as existing satellite ontologies.

4. The conceptualization of *offenders*

The remaining candidate terms obtained, not being already present in FunGramKB, can be considered for further enrichment and expansion of the conceptual repository of this knowledge base. In order to show the process of conceptual modeling and hierarchization in the taxonomy, four lexical units have been selected for this paper: a) "felon;" b) "infringer; c) "misappropriator;" and d) "impersonator." All these words share the "+human" and "+agent" semantic properties, which highlight a central concern in criminal law, that is, "who" is responsible for unlawful acts. As will be shown below, the four terms belong to the specialized legal domain, and not to general knowledge, and all of them refer to criminals that commit an offense against non-material goods, such as intellectual property, secrets, personal information or identity.

4.1 *Felon*

The term *felon* appears in some renowned general dictionaries of English language, including the Longman Dictionary of Contemporary English (LDCE), the Oxford Advanced Learner's Dictionary (OALD), and the Cambridge Dictionary (CD). In the latter, for example, "felon" is defined as "a person who is guilty of a serious crime" and is tagged as belonging to the field of Law. This information, specifically the semantic tag, makes this term a potential concept for the Satellite Ontology. In addition, the definition of "felon" in the Black's Law Dictionary (BLD)—one of the most authoritative specialized dictionaries in the field of Law—is even more specific: "a person who has been convicted of a felony", and "felony" is in turn explicitly labeled as a term belonging to the field of criminal law. Finally, according to all the specialized sources consulted, "felony" includes serious crimes, such as burglary, arson, rape or murder. Thus, in the conceptual organization of this term, a number of subordinate concepts have to be created, namely "burglar", "arsonist" and so forth, as "felon" becomes an umbrella concept for all types of criminals who commit felonies.

The translation of terminological units into conceptual units to populate the ontology in FunGramKB is done using the COREL representation language (Periñán-Pascual & Mairal-Usón, 2010). In this regard, and following the FunGramKB notation system, the concept "felon" must be formalized as +FELON_00, and its conceptual path in the ontology is as follows: #ENTITY > #PHYSICAL > #OBJECT > #SELF_CONNECTED_OBJECT > +HUMAN_00 > +WRONGDOER_00 > +CRIMINAL_00 > +FELON_00. "Felon" belongs to the #ENTITY subontology, whose concepts are linguistically realized by nouns, and to the following metacognitive dimensions: #PHYSICAL, #OBJECT and #SELF_CONNECTED_OBJECT. Its superordinate concept is +CRIMINAL_00, thus inhering all the conceptual properties of this concept; in other words, the meaning of "felon" is subsumed by +HUMAN_00, +WRONGDOER_00 and +CRIMINAL_00, i.e., a felon is a human criminal. It is important to highlight that +FELON_00 will be a superordinate concept and, thus, used in the conceptual definition of the terminal concepts $BURGLAR_00 or $ARSONIST_00.

The MP constitutes the most important semantic information of a concept in FunGramKB and it expresses non-inherited specific information about the entity. Below is the MP of +FELON_00:

+(e1: +BE_00 (x1: +FELON_00)Theme (x2: +CRIMINAL_00)Referent)

+(e2: +DO_00 (x1)Theme (x3: +FELONY_00)Referent)

Its natural language equivalent is "felon is a criminal who commits a felony." It is relevant to mention that the concept +FELONY_00 was already part of the Satellite Ontology of organized crime and terrorism and, consequently, the MP of +FELON_00 is conceptually linked to the semantic traits of "felony," being a serious crime punishable with imprisonment or death, as can be seen in its MP:

+(e1: +BE_00 (x1: +FELONY_00)Theme (x2: +CRIME_00)Referent)

+(e2: +BE_01 (x2)Theme (x3: +SERIOUS_00)Attribute)

*((e3: +DO_00 (x1)Theme (x4: +PUNISHMENT_00)Referent)(e4: +BE_00 (x4)Theme (x5: +PRISON_00 ^ +DEATH_00)Referent))

This inference process, where conceptual information from one concept is linked to a different concept outside a purely inheritance process, reduces redundancy and enriches the informative capacity of the ontology, thus allowing in the case of +FELON_00 the creation of a formally short conceptual definition that is nonetheless rich and expanded with the reference to the concept +FELONY_00. Finally, since "felon" is part of a lay-person's knowledge, as it is included in general dictionaries, a concept should also be created to populate FunGramKB's Core Ontology, but as a terminal concept: $FELON_00. Consequently, the basic concept +FELON_00 in the Satellite Ontology of criminal law becomes a "mirror" concept of the terminal concept $FELON_00 in the Core Ontology.

4.2 *Infringer*

Consultation of lexicographical sources for the term *infringer* showed that it is not included in general dictionaries (LDCE, OALD and CD). However, it is listed as an entry in specialized sources such as BLD and Merriam-Webster Law Dictionary (MWLD). "Infringer" is, thus, a concept that does not belong to an average person's knowledge and, as such, there is no place for it in a general ontology. However, being acknowledged as specialized knowledge in the legal field, it is a candidate concept to populate a specialized repository. It is for this reason that its conceptualization in the Satellite Ontology in FunGramKB seems pertinent, more precisely as the basic concept +INFRINGER_00.

Both BLD and MWLD define this term as a person who violates the rights of another person, being those rights either patents, copyrights, or trade secrets. In other words, it refers to a criminal that acts against the intellectual property of another person. For its conceptualization in FunGramKB, similarly to the previous concepts presented above, +INFRINGER_00 belongs to the #ENTITY subontology, whose concepts are linguistically realized by nouns. Likewise, in the hierarchical organization of the ontology, its superordinate is +CRIMINAL_00, thus inheriting all the conceptual properties of this concept. On the other hand, the non-inherited specific conceptual information is represented in the MP using the COREL notation system as follows:

+(e1: +BE_00 (x1: +INFRINGER_00)Theme (x2: +CRIMINAL_00)Referent)

+((e2: +USE_00 (x1)Theme (x3:+PROPERTY_00)Referent (f1: $LEGAL_N_00)Manner) (e3: +BE_00 (x3)Theme (x4: +THOUGHT_00 | +CREATION_00)Referent))

The first predication in this MP identifies an infringer as a criminal. The second predication expresses that the infringer uses a property in an illegal manner, further specifying in the third predication that the property is intellectual, rather than physical or material. The second and the third predications are cognitively bounded together since the former needs the specification of the latter in order to complete its meaning. The syntax of COREL represents this conceptual binding by enclosing the two predications between parentheses (Jiménez Briones & Luzondo Oyón, 2011: 31).

4.3 *Misappropriator*

Misappropriation is extensively illustrated in manuals and reference works in the legal field, such as Steinberg (2021) or Flinn (2022). As for lexicographical resources, “misappropriator” does not appear in the general dictionaries, although “misappropriation does. As with “felon,” it is the BLD which offers the most comprehensive definition of this term (BLD, 1088):

> 1. The application of another's property or money dishonestly to one's own use [...]
> 2. *Intellectual property*. The common law tort of using the noncopyrightable information or ideas that an organization collects and disseminates for a profit to compete unfairly against that organization, or copying a work whose creator has not yet claimed or been granted exclusive rights in the work.

As we can see, there are two related but differentiated senses of the word. The first sense is more generic and refers to the act of taking something in an improper manner. The second sense is more specific and refers to the act of using information from a second agent (for example a company) in a fraudulent way with the result of unfair competence. The term “embezzlement” is also related to this state of affairs, but “misappropriator” is specifically related to “intellectual property”. We will base our COREL definition below mainly on BLD’s definition for the second sense above:

+(e1: +BE_00 (x1: +MISAPPROPRIATOR_00)Theme (x2:+CRIMINAL_00)Referent)

+(e2: +USE_00 (x1)Theme (x3:+INFORMATION_00 | +THOUGHT_00)Referent

(f1: (e3: past n +HAVE_00 (x1)Theme (x3)Referent))Condition (f2: (e4:

+COMPETE_00 (x1)Theme (x4: +COMPANY_00)Referent))Purpose (f3: $LEGAL_N_00)Manner)

The second predication could be roughly translated as “someone uses information or ideas, which do not belong to that person, with some illegal purpose.” In this case the meaning of “belong” or “own” is expressed conceptually with the generic event +HAVE_00. We additionally introduce the negative operator (“n”) and past operator (“past”) to emphasize that the information being used was not previously owned by the agent. Also in the second predication, there are three satellite participants (identified with “f”) that represent peripheral information, as opposed to the core participants in the predications that are identified with “x”. The satellites in this MP indicate condition, purpose and manner. The specification of the second satellite (f2) is a predication. It introduces the concept +COMPETE_00, which is already present in the Core Ontology with its two main components: Thematic Frame (TF) and MP. First, the TF is a general statement that appears in the definition of all the events (i.e., verbs) in FunGramKB’s ontology. It establishes the necessary participants that take part in the state of affairs evoked. In our example, the TF includes both “humans” or “animals” as potential actors of competing, but they are connected with an “or” operator (^) which in our case prevents any connection between the act of misappropriation with an animal acting as the agent of that action. In fact, in the MP of +MISAPPROPRIATOR, the Theme is identified with the misappropriator (x1) and the Referent is a company (x4). In the hierarchical organization of FunGramKB, +COMPANY_00 is subsumed to +ORGANIZATION_00 > +PEOPLE_00, hence a human participant. Second, the MP of +COMPETE_00 refers to the fact that two different agents try to obtain the same object or situation, here generally defined as a “Referent:”

THEMATIC FRAME:

(x1: +HUMAN_00 ^ +ANIMAL_00)Theme

(x2: +HUMAN_00 ^ +ANIMAL_00)Referent

MP:

+((e1: +TRY_00 (x1)Theme (x3: (e2: +OBTAIN_00 (x1)Theme (x4)Referent))Referent)(e3: +TRY_00 (x2)Theme (x5: (e4: +OBTAIN_00 (x2)Theme (x3)Referent))Referent))

4.4 *Impersonator*

The last concept we have chosen to illustrate the conceptual modeling of legal concepts refers to an agent who commits a crime of the type of violation of property rights. As in the previous cases, the first step consists in analyzing lexicographic sources in order to identify the semantic particularities of this term. The CALD does register this word with a basic definition, without delving into how the criminal process develops. Furthermore, this dictionary (and similarly the OED and Longman Dictionary) does not include any criminal or legal aspect of the term: “someone who copies the way another person looks and behaves, usually in order to entertain people.” On the other hand, specialized dictionaries offer many more specific semantic aspects for conceptualization. For instance, BLD, which does not include “impersonator,” but “impersonation,” defines the latter as follows:

> (18c) The act of impersonating someone. Also termed personation. false impersonation. (1878) The crime of falsely representing oneself as another person, usu. a law-enforcement officer, for the purpose of deceiving someone. See 18 USCA §§ 912-917. Also termed false personation. (Cases: False Personation ~~ 1.]

Similarly, the Oxford Dictionary of Law (ODL) also defines the term “impersonation” and it adds some circumstances in which this type of crime may take place:

> n. Pretending to be another person. It is an offence to impersonate a woman's husband in order to persuade her to have sexual intercourse (rape), to impersonate the holder of a Crown office in order to gain access to prohibited places, and to impersonate a police officer, a variety of public officials, a voter, or a juror. Obtaining property, services, or certain financial advantages through impersonation may amount to a crime of deception.

These definitions need to be synthesized in order to be translated into COREL. In doing so, we will focus on the main state of affairs involved in this type of crime and we will not include in our definition the examples of “impersonation” provided by the dictionaries:

+(e1: +BE_00 (x1: +IMPERSONATOR_00)Theme (x2: +CRIMINAL_00)Referent)

+(e2: +STEAL_00 (x1)Theme (x3: $FEATURE_00)Referent (f1:

+HUMAN_00)Origin (f2: (e3: +DECEIVE_00 (x1)Theme (x4)Referent)Purpose)))

In this definition, similarly to the previous ones, the first predication identifies an impersonator as a criminal agent, while the second predication includes its identifying characteristics, that is, the specific act that this agent commits: he/she steals some characteristics ("Features") that are representative of another person ("Origin") with the purpose of deceiving someone else.

5. Conclusions

This paper has shown how DEXTER, an online tool for terminology extraction, can be used in the ontological enrichment of knowledge bases. As we have tried to show, DEXTER offers many advantages to both terminologists and scholars interested in the enrichment of specialized vocabulary resources. This tool's statistical engine has proven to have a great capacity of retrieval of terminology units as well as a high percentage of accuracy. In addition, DEXTER, which is open online for any user, offers multiple options for managing word lists and filter non-relevant units. In the second part of the paper, we have tried to connect the terminological-extraction work with a conceptual modeling task, i.e., the process whereby terminological units are formalized as specialized conceptual units. As demonstrated in this paper, concepts defined semantically with a formalization language can become part of FunGramKB, a knowledge base which contains both ontological and linguistic information and which can be used for NLP tasks. As a representative example of the conceptualization phase, we have analyzed the semantics of a specific set of lexical units from the domain of criminal law. The set is related to "offenders," that is, people who commit specific crimes of varying severity, such as "felon," "impersonator," "misappropriator" or "infringer." While the two later units can be considered candidates to populate a specialized module within FunGramKB's ontological component, since they are essentially terminological, "felon" must be considered to have a dual conceptual nature. In this case, dictionaries show that it is a concept belonging to the knowledge of an average speaker and, at the same time, a concept to which experts add specialized semantic information. As a result, the concept "felon" should appear in the general domain ontology in FunGramKB as a terminal concept, while a second instance of the concept should mirror in the corresponding specialized ontology within the knowledge base. Thus, the present paper with this selection of examples belonging to the field of criminal law has shown the process for the incorporation of terminological units retrieved from a specialized text collection using this term-extraction tool that has hence revealed great potential and made it possible for the development of a specialized ontology to cover the full and broader field of criminal law.

Acknowledgments

This article is based on the R&D Project within the framework of the "Programa Operativo FEDER Andalucía 2014-2020", code B-HUM177-UGR18. This work was also funded by MCIN/AEI/10.13039/501100011033 under grant PID2020-112827GB-I00.

References

Alameda-Hernández, Á. & Felices-Lago, Á. (2016). The integration of the concept +CRIME_00 in FunGramKB and the conceptualization or hierarchization problems involved. In C. Periñán-Pascual & E. Mestre-Mestre (Eds.), *Understanding Meaning and Knowledge Representation: From Theoretical and Cognitive Linguistics to Natural Language Processing* (pp. 319-340). Newcastle: Cambridge Scholars Publishing.

Alameda-Hernández, Á. & Felices-Lago, Á. (2017). Crimes against children: an apparently terminological knowledge representation of entities in FunGramKB. *Journal of Computer-Assisted Linguistic Research*, 1, 1-19. doi: 10.4995/jclr.2017.7785

Bárcena, E., Read, T. & Arús Hita, J. (Eds). (2014). *Languages for Specific Purposes in the Digital Era*. London: Springer.

Breuker, J., Casanovas, P., Klein, M. & Francesconi, E. (2009). *Law, Ontologies and the Semantic Web. Channelling the Legal Information Flood*. Amsterdam: IOS Press.

Cabré, M. T. (1999). *Terminology: Theory, Methods, and Applications.* Amsterdam and Philadelphia: John Benjamins.

Cambridge Dictionary. Retrieved from https://dictionary.cambridge.org/es/diccionario/ingles/

Carrión Delgado, G. (2012). Extracción y análisis de unidades léxico-conceptuales del dominio jurídico: un acercamiento metodológico desde FunGramKB. *Revista Electrónica de Lingüística Aplicada*, 11, 25-39.

Carrión-Delgado, G. & Felices-Lago, Á. (2014). La jerarquización cognitiva de las entidades en la ontología satélite del crimen organizado y el terrorismo en FunGramKB. In C. Vargas Sierra (Ed.), *TIC, trabajo colaborativo e interacción en Terminología y Traducción. Colección Interlingua, vol. 132* (pp. 591-608). Granada: Comares.

Casanovas, P., Sartor, G., Biasiotti, M.A. & Fernández Barrera, M. (2011). Theory and methodology in legal ontology engineering: experiences and future directions. In G. Sartor, P. Casanovas, M. A. Biasiotti & M. Fernández-Barrera (Eds.), *Approaches to Legal Ontologies, Theories, Domains, Methodologies* (pp. 3-14). Berlin: Springer.

Chu, W. W. (2014). Data mining and knowledge discovery for Big Data. *Studies in Big Data*, 1, 153-192.

Felices-Lago, Á. & Ureña Gómez-Moreno, P. (2012). Fundamentos de la creación subontológica en FunGramKB. *Onomázein*, 26(2), 49-67. doi: 10.7764/onomazein.26.02

Felices-Lago, Á. & Ureña Gómez-Moreno, P. (2014). FunGramKB Term Extractor: a key instrument for building a satellite ontology based on a specialized corpus. In B. Nolan & C. Periñán (Eds.), *Language Processing and Grammars: The Role of Functionally Oriented Computational Models. (Studies in Language Series)* (pp. 251-269). Amsterdam and Philadelphia: John Benjamins.

Felices-Lago, Á. (2015). Foundational considerations for the development of the Globalcrimeterm subontology: a research project based on FunGramKB. *Onomázein* 31, 127-144. doi: 10.7764/onomazein.31.9

Felices-Lago, Á. & Alameda-Hernández, A. (2017). The process of building the upper-level hierarchy for the aircraft structure ontology to be integrated in FunGramKB. *Revista de Lenguas para Fines Específicos,* 23, 86-110.

Felices-Lago, A. & Ureña Gómez-Moreno, P. (2020). Conceptualización de entidades terminológicas en una subontología de derecho penal. Análisis del concepto superordinado +DRUG_00 en FunGramKB. *Revista de Lingüística y Lenguas Aplicadas,* 15(1), 15-25. doi: 10.4995/rlyla.2020.12772

Flinn, P. J. (2000). *Handbook of Intellectual Property Claims and Remedies*. New York: Wolters Kluwer.

Garner, B. (Ed.). (2009). *Black's Law Dictionary*. Saint Paul: Thomson Reuters.

Ikonomakis, E., Kotsiantis, S. Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, 4(8), 966-974.

Jiménez Briones, R. & Luzondo Oyón, A. (2011). Building ontological meaning in a lexico-conceptual knowledge base. *Onomázein* 23, 11-40.

Lausch, A., Schmidt, A. & Tischendorf, L. (2015). Data mining and linked open data. New perspectives for data analysis in environmental research. *Ecological Modelling*, 295, 5-17. 10.1016/j.ecolmodel.2014.09.018

Liebwald, D. (2007). Semantic spaces and multilingualism in the law: the challenge of legal knowledge management. In P. Casanovas, M. A. Biasiotti, E. Francesconi & M. T. Sagri (Eds.), *Proceedings of LOAIT* (pp. 131-148). Stanford: Stanford University.

Longman Dictionary of Contemporary English (2010). Harlow: Pearson Longman.

Merrian-Webster's Dictionary of Law (1996). Springfield, Mass: Merriam-Webster. Retrieved from https://www.merriam-webster.com/legal

Oxford Advanced Learner's Dictionary (2020). Oxford: Oxford University Press.

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

Periñan-Pascual, C. (2018). DEXTER: A workbench for automatic term extraction with specialized corpora. *Natural Language Engineering*, 24(2), 163-198. doi: 10.1017/S1351324917000365

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

Periñán-Pascual, C. & Arcas-Túnez, F. (2010). The architecture of FunGramKB. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 2667-2674). Valletta, Malta: European Language Resources Association (ELRA).

Periñán-Pascual, C. & Arcas-Túnez, F. (2014). La ingeniería del conocimiento en el dominio legal: la construcción de una Ontología Satélite en FunGramKB. *Signos: Estudios de Lingüística,* 47(84), 113-139. doi: 10.4067/S0718-09342014000100006

Periñán-Pascual, C. & Mairal-Usón, R. (2010). The COREL grammar: a conceptual representation language. *Onomázein*, 21, 11-45.

Rajni, J., Ruchika, M. & Abha. J. (2015). Techniques for text classification: literature review and current trends. *Webology*, 12(2). Article 139. Retrieved from https://www.webology.org/2015/v12n2/a139.pdf

Samaha, J. (2011). *Criminal Law*. Wadsworth: Cengage Learning.

San Martín, A. & Faber, P. (2014). Deep semantic representation in a domain-specific ontology: linking EcoLexicon to FunGramKB. In B. Nolan & C. Periñán-Pascual (Eds.), *Language Processing and Grammars* (pp.271-296). Amsterdam and Philadelphia: John Benjamins.

Steinberg, M. I. (2021). *Securities Regulation: Liabilities and Remedies*. Law Journal Press.

Ureña Gómez-Moreno, P. (2016). La lucha contra el terrorismo y la delincuencia organizada: una visión desde la lingüística y la ingeniería del conocimiento. *Miscelánea*: *A Journal of English and American Studies* 53, 107-123.

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