ONOMÁZEIN
Revista semestral de lingüística, filología y traducción

PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE
FACULTAD DE LETRAS

Foundational considerations for the development of the *Globalcrimeterm* subontology: A research project based on FunGramKB

Ángel Felices Lago
Universidad de Granada
España

ONOMÁZEIN 31 (junio de 2015): 127-144
DOI: 10.7764/onomazein.31.9

Ángel Felices Lago: Departamento de Filologías Inglesa y Alemana, Facultad de Ciencias Económicas y Empresariales, Universidad de Granada, España. | Correo electrónico: afelices@ugr.es

Fecha de recepción: marzo de 2014
Fecha de aceptación: diciembre de 2014

31
Junio
2015

Abstract

This paper describes the theoretical foundations, the general methodological guidelines and the specific tasks for the development of the *Globalcrimeterm* project, a domain-specific subontology based on a specific area of criminal law (international cooperation against terrorism and organized crime) within the architecture of FunGramKB, which is a multipurpose lexico-conceptual knowledge base for natural language processing (NLP) systems. One of the features of this subontology is, firstly, its commitment to structure its concepts under the postulates of deep semantics, unlike the more traditional approach only oriented towards surface semantics, and, secondly, to orientate the tasks of terminologists and knowledge engineers who wish to expand the general knowledge of FunGramKB Core Ontology and, at the same time, integrate the specialized knowledge through its representation in a domain-specific subontology such as *Globalcrimeterm*.

Keywords: FunGramKB; legal ontology; terminology; *Globalcrimeterm*; NLP.

1. Introductory remarks[1]

Nowadays modern society has access to an unprecedented volume of information thanks to the Internet and an increasing number of sources of knowledge. Paradoxically enough, access to such a wealth of information is often complicated and time-consuming, partly because of the intricacies involved in finding precise information easily and accurately. The latest developments in the field of Artificial Intelligence and the Semantic Web highlight the need to design intelligent systems capable of processing human queries in a natural language as well as retrieving only the relevant information matching a specific set of input questions. A substantial yet initial step in this direction has been taken with the development of semantic mark-up metalanguages such as OWL (Web Ontology Language) and RDF (Resource Description Framework), which are currently being applied with a view to endowing the web with superficial semantic knowledge and to facilitating the retrieval of data for web users. Despite all the numerous attempts to build a more accessible and intelligent web, there is still a strong need for a truly efficient model of communication between humans and machines, and the way the former interact with the latter to access the required information. We need a more robust model of semantic representation to provide machines with the capacity to understand human language and an organised conceptual system to also allow them to mimic the understanding of the human world.

The guidelines of the research project presented in this article are part of an ambitious scientific programme carried out by an international group of researchers from various universities. In general terms, the proposal I intend to put forward is aimed at contributing with specific resources to the implementation of innovative solutions (at a later stage) in the field of human-machine communication and the application of natural language processing techniques to a broad range of human activities. The initial steps taken in this large-scale research framework originated in 2004 with the seminal work by Periñán and Arcas, who explained and illustrated the development of the groundbreaking system called FunGramKB, an advanced multipurpose lexical conceptual knowledge base for natural language processing (NLP) systems (Periñán & Arcas, 2004, 2005, 2006, 2007). Ever since its creation, FunGramKB has been continuously developing to meet the goals of cutting-edge artificial intelligence research as well as to cover some of the long-standing yet unresolved issues faced by NLP, especially those concerning the aforementioned problems with information access and human-machine communication. The scope of application of FunGramKB is promising and concerns many relevant areas of NLP, including dialogue systems, expert agents, data mining or information retrieval, to name but a few.

This knowledge base has also proved to be a rich explanatory framework where a broad meaning construction model of language such as the Lexical Constructional Model (Mairal & Ruiz de Mendoza, 2009; Ruiz de Mendoza & Mairal, 2008, 2011) can be anchored. Consequently, the expression 'knowledge base' is defined here as a computational repository in which conceptual and linguistic knowledge is stored (and accessed) in a connected, meaningful, and efficient way.

Before moving on to an explanation as to how the project presented here actually relates to the general framework of FunGramKB, it is first necessary both to formulate the main objective of this research and to review some

[1] This article is based on research carried out within the framework of the Project FFI2014-53788-C3-1-P, which is funded by the Spanish Ministry of Economy and Competitiveness. It is also based on the previous Project code no. FFI2010-15983 and entitled "Development of a subontology in a multilingual context (English, Spanish and Italian): Using FunGramKB in the field of international cooperation in criminal matters: Terrorism and organised crime".

of the main theoretical principles that have inspired the whole process in the diverse scientific areas involved, due to its multidisciplinary nature. Then, a set of methodological assumptions and procedural details will be established in order to carry out the relevant tasks for the implementation of the specific objectives of the project, which can be used as a guideline for the construction of similar subontologies based on a deep semantics.

2. Main objective of the *Globalcrimeterm* project

The main objective of this project is the construction of a terminological subontology structuring its concepts under the postulates of deep semantics, unlike the more traditional approach only oriented towards surface semantics. More specifically, this project revolves around the design and implementation of this subontology, together with the population of this module and its three respective terminological lexicons (English, Italian and Spanish).

The resultant ontology might be presented as a hierarchical network according to the meaning postulates of FunGramKB's knowledge base. Furthermore, its interaction within the already existent general ontology would be allowed in order to apply it to comprehension tasks of a natural language. The subsequent repository obtained on the chosen topic (criminal law: terrorism and organised crime) could be used by both humans (through a dictionary-like interface) and by machines (through its future application to natural language processing (NLP) systems). FunGramKB has already proved its multifunctional character in other applications and former projects funded by the Spanish Ministry of Economy and Competitiveness (e.g. DGI[MICINN]: FFI2010-17610; FFI2008-05035-C02-01 or MEC: HUM2005-02870)[2] as well as its potentially reusable character, which will be fully verifiable when applied to two main NLP systems: automatic translation and retrieval of information through future complementary projects. Indeed, the aim is to go beyond the elaboration of a mere data base of terminology on a topic of the upmost importance and relevance on an international scope, and to also provide an intelligent system which is able to automatically relate concepts, terms and written material from different sources through an editor and a web browser.

3. Project foundations

This project has its roots in several scientific fields of knowledge which it endeavours to relate to one another and obtain the best possible output from their coexistence. On the one hand, it deals with the cognitive and functional approaches to language, mainly in some aspects of *Role and Reference Grammar* (Van Valin & LaPolla, 1997; Van Valin, 2005) and Mairal & Ruiz de Mendoza's (2008, 2009) *Lexical Constructional Model* (see www.lexicom.es)[3]; on the other hand, it deals with the potential of the Lexical Constructional Model in the field of NLP and, in particular, with its semantic *web*. For this it is necessary to begin with the influential work mentioned above and carried out by Periñán & Arcas (2004, 2005, 2006, 2007), who have developed a lexical conceptual knowledge base, namely *FunGramKB* (www.fungramkb.com). And finally, taking into account the development of general ontologies (Gruber, 1993, or McGuinness et al., 2000), ontologies used in legal fields (Breuker et al., 2005; Valente, 2005; Breuker et al., 2008, and Sartor et al., 2011), the precedents of onomasiological lexicology and lexicography pointed out by Martín-Mingorance (1994) and the axiological aspects which criminal law involves (Felices, 2010), this project proposes

[2] DGI: Directorate-General for Research; MICINN: Former Ministry of Science and Innovation from Spain; MEC: Former Ministry of Education and Science from Spain.

[3] For other project's precedents see Simon C. Dik's Functional Grammar (1989) and Martín-Mingorance's Functional-Lexematic Model (1998).

a novel, multipurpose and interactive application which takes as its starting point its concept of deep semantics.

3.1. Linguistic level

Role and Reference Grammar (RRG) adopts a cognitive and communicative perspective of language; that is to say that grammatical rules and morphosyntactic structures should be explained in terms of their communicative and semantic functions. Van Valin and LaPolla's RRG is a monostratic theory given that syntactic and semantic components are effectively expressed without the need to use abstract syntactic representations. Consequently, the aforementioned syntactic and semantic components are projected directly, following an algorithm link which includes a set of rules which facilitate the syntactic-semantic interface. Furthermore, RRG contains three fundamental levels of representation: (i) one which captures the meaning of linguistic expressions according to an inventory of logical structures; (ii) one which represents the syntactic structure of the clause based on universally valid distinctions; and (iii) one which represents the structure of the information of a speech act.

On the other hand, Mairal and Ruiz de Mendoza's *Lexical Constructional Model* (LCM) offers a new model which attempts to explain all the aspects implied in the construction of meaning, including those which go beyond purely grammatical ones such as the pragmatic (implicational) (level 2), illocutionary (level 3) and discourse (level 4) levels. Thus, its final result is a comprehensive representation of all those aspects which the meaning of a statement involves. Therefore, LCM offers a number of principles, axioms, labels, etc., which give rise to a whole model of representation of meaning which could fulfil the need of developing the semantic *web*.

Both the RRG and the LCM share two features which are fundamental for a computational model of language:

- A functionalist vision of the language which allows us to capture syntactic-semantic generalizations which are fundamental to explaining the semantic motivation of grammatical phenomena.
- A strong commitment regarding the typological adequacy involved in universal distinctions as an essential part of the linguistic framework. Typological adaptation is a *conditio sine qua non* in multilinguistic models.

3.2. Natural Language Processing and FunGramKB

The latest developments in the field of Artificial Intelligence and the semantic web point towards the designing of "intelligent agents" which are capable of processing consultations made in a natural language. With the aim of facilitating the retrieval and extraction of information from the World Wide Web in an intelligent way, languages with semantic labels have been invented such as OWL (*Web Ontology Language*). Even if their results are promising, they require a model of representation with a strong semantic foundation which is capable of producing linguistic labels with full meaning and which is machine-usable so that a machine is able to understand the consultation made in a natural language and to retrieve the information required. Furthermore, the problem of the automatic filter when searching for relevant texts is exponentially complicated in multilinguistic contexts (Aguado de Cea et al., 2007; Mairal & Ruiz de Mendoza, 2009; Montiel et al., 2007; Periñán & Arcas, 2007, 2008; Periñán & Mairal, 2009). It is within this context that the development of the knowledge base FunGramKB is elaborated. Indeed, it allows the enrichment of the applications for NLP: e.g. intelligent agents for the processing of information, prototypes of automatic translation or dictionaries based on conceptual searches, which could be catalogued as "dictionaries of the third millennium" (Periñán & Arcas, 2006).

This orientation towards computing factors introduces substantial modifications in the linguistic model under scrutiny, which ceases to have a lexicalist foundation and has taken on a conceptualist or ontological stance. Moreover, some of the most remarkable advantages of using this conceptualist approach are mentioned, such as its greater expressivity in representation and its access to encyclopedic knowledge, both unapproachable from a merely lexical standpoint. For reasons of space, a detailed vision of all its components cannot be given but can be consulted in Periñán & Arcas (2004, 2005, 2007), Periñán & Mairal (2009) and in Mairal & Periñán (2009, 2010).

If one abides by the distinction made by Velardi et al. (1991) between surface semantics and deep semantics, it is predictable that one of the consequences of using FunGramKB as regards knowledge bases such as SIMPLE or *EuroWordnet* is that the stance adopted is that of conceptual representation in deep semantics. Why is this approach more viable?

Computer systems with surface semantics in their knowledge base have information about lexical relations which are established between lexical units. In other words, the representation of a word's meaning is made exclusively through specifying the relation that that word has with others. This is the case, for example, of *WordNet*, which although possessing defining texts for *synsets*, this information is not machine-usable, subsequently leaving the relations between *synsets* as the only option. Although it is easy, and above all, quick, to populate a knowledge base in this way, difficulties are found when one attempts to apply this method to the representation of conceptual units such as REMEMBER, FORGET, LOVE, etc., which one finds difficult to express in terms of relational meaning. What is more, knowledge bases with deep semantics such as FunGramKB develop a language of representation (or interlingua), namely COREL (*Conceptual Representation Language*), which allows us to define all conceptual units with the added advantage that conceptual relations can be equally obtained through mechanisms of inheritance or inference on the meaning postulates (cf. Periñán & Arcas, 2005).

Hence, if one of our aims is to represent and manage knowledge in one application we should specify the modules or components which make up the format of this application. In this sense, Periñán & Arcas (2007, 2010) and Periñán & Mairal (2011: 16-18) distinguish linguistic and non-linguistic information, including both three major knowledge levels, consisting in turn of several independent but interrelated modules:

a) The linguistic level (linguistic knowledge):

1) Lexical level:

- The *Lexicon* stores morphosyntactic, pragmatic and collocational information about lexical units.
- *Morphicon* handles cases of inflectional morphology.

2) Grammatical level:

- *Grammaticon* stores the constructional schemata which help Role and Reference Grammar to construct the semantics-to-syntax linking algorithm (Van Valin & LaPolla, 1997; Van Valin, 2005). The Grammaticon is composed of several *Constructicon* modules that are inspired in the four levels of meaning construction formulated in the LCM:

(i) an argument structure layer, which contains Conceptual Logical Structures (CLSs) and argument structure constructions;

(ii) an implicational level, with constructional configurations, based on low-level situational models (or scenarios), which contain fixed and variable elements where the default meaning interpretation carries a heavily conventionalized implication;

(iii) an illocutionary level, which features illocutionary constructions, with fixed and variable elements based on high-level situational models;

(iv) a discourse level, which deals with cohesion and coherence phenomena from the point of view of the activity of discourse constructions based on high-level non-situational cognitive models like reason-result, cause-effect or condition consequence.

b) The conceptual level (non-linguistic knowledge):[4]

- *Ontology* is presented as a hierarchical catalogue of the concepts that a person has in mind, so here is where semantic knowledge[5] is stored in the form of meaning postulates. The ontology consists of a general-purpose module (i.e. Core Ontology) and several domain-specific terminological modules (i.e. satellite ontologies or subontologies).

- *Cognicon* stores procedural knowledge by means of scripts, that is, conceptual schemata in which a sequence of stereotypical actions is organised on the basis of temporal continuity, and more particularly on Allen's temporal model (Allen, 1983; Allen & Ferguson, 1994); e.g. 'dine in a restaurant', 'celebrate a wedding', or 'launder money', etc.

- *Onomasticon* stores information about instances of entities and events such as *Big Ben*, *September 11*, *Osama Bin Laden*, *Leaving Las Vegas*, etc.: episodic knowledge. This module stores two different types of schemata (i.e. snapshots and stories), since instances can be portrayed synchronically or diachronically.

Here only the basic ontology and developments of terminological subontologies are referred to, although it is necessary to mention, at least, that the other two cognitive modules—the cognicon and the onomasticon—are expressed using the same language, COREL, and the same conceptual units employed in the ontology. In this sense, three concepts are dealt with (Periñán & Arcas, 2007; Mairal & Periñán, 2009; Periñán & Mairal, 2011):

(i) *Metaconcepts*, preceded by symbol # (e.g. #ABSTRACT, #COMMUNICATION, #MATERIAL, #PHYSICAL, #PSYCHOLOGICAL, #QUANTITATIVE, #SOCIAL, etc.), constitute the upper level in the taxonomy. The result amounts to forty-two metaconcepts distributed in three subontologies: #ENTITY, #EVENT and #QUALITY.

(ii) *Basic concepts*[6], preceded by symbol + (e.g. +VIOLENT_00, +CRUEL_00, +CRIME_00, +TRIAL_00, +OFFEND_00, +PUNISH_00, +MURDER_00, etc.), are used in *FunGramKB* as defining units which enable the construction of meaning postulates for basic concepts and terminals, as well as taking part as selectional preferences in thematic frames.

(iii) *Terminals* (e.g. $ASSASSINATION_00, $FELONY_00, $GANGSTER_00, $_00, $CONSPIRE_00, $DISHONEST_N_00, etc.) are headed by the symbol $. The borderline between basic concepts and terminals is based on their definitory potential to take part in meaning postulates. Hierarchical structuring of the terminal level is practically non-existent.

[4] This Project concentrates its activity at this level, particularly on the ontology, but research on the cognicon and onomasticon is also being developed.
[5] The underlined types of knowledge follow the distinctions established within the framework of cognitive psychology.
[6] The examples of basic and terminal concepts indicated here have been obtained from the *Globalcrimeterm* subontology under construction.

Basic and terminal concepts in FunGramKB are provided with semantic properties which are captured by *thematic frames* and *meaning postulates*. Every event in the ontology is assigned one single thematic frame, i.e. a conceptual construct which states the number and type of participants involved in the prototypical cognitive situation portrayed by the event (Periñán & Arcas, 2007)[7]. Moreover, a meaning postulate is a set of one or more logically connected predications (e_1, e_2, ... e_n), i.e. conceptual constructs that represent the generic features of concepts. As stated above, the basic concepts are the main building blocks of these types of constructs in the Core Ontology.

Referring back to the linguistic and non-linguistic knowledge, it is important to note that the cognitive level covers all those properties which are universal, that is to say, common to all languages, the lexical level covers the description of the idiosyncratic properties of each language. From a metatheorical stance, the inclusion of a knowledge base such as FunGramKB introduced a far reaching change in linguistic theory as the model no longer begins with the lexical component but with the conceptual level. Consequently, the lexical component ceases to be the starting engine of the linguistic machinery in order to be the recipient of a wealth of information which its conceptual meaning gives it, and more specifically, which the ontology gives it. Figure 1 (source: www.fungramkb.com/) represents this cognitive turn, where hypothetically we would place ourselves on the upper part and postulate a conceptual level which feeds the different lexica of each language. In essence, the weight of the semantic description lies in the ontology, whereas the lexical entries remain extremely simplified, albeit with a high degree of expressivity and linguistic information, where semantic knowledge can be inferred with the help of the reasoner.

As mentioned above, FunGramKB, unlike other applications, defines each one of the concepts. For this, it uses COREL, which like other languages, possesses its own semantics and syntax. Let us consider the meaning postulate of the concept $COCAINE_00[8]:

+(e1: +BE_00(x1: +COCAINE_00)Theme (x2: +DRUG_00)Referent)

*((e2: +MAKE_00 (x3)Theme (x1)Referent (x4: $COCA_00)Means)

*((e3: +BE_00 (x4)Theme (x5: +LEAF_00) Means)

*(e4: +GIVE_00 (x1)Agent (x6: +PLEASURE_00 & m+ENERGY_00)Theme (x1)Origin (x7)Goal)

*(e5: +TAKE_00 (x1)Theme (x2)Referent (x3:+NOSE_00)Means)

*(e6: +INGEST_00 | +SELL_00 | +BUY_00 (x8) Agent (x1)Theme (x6)Origin (x7)Goal (x9: +LEGAL_N_00) Attribute)

An approximate translation of the meaning postulate to a natural language would be the following:

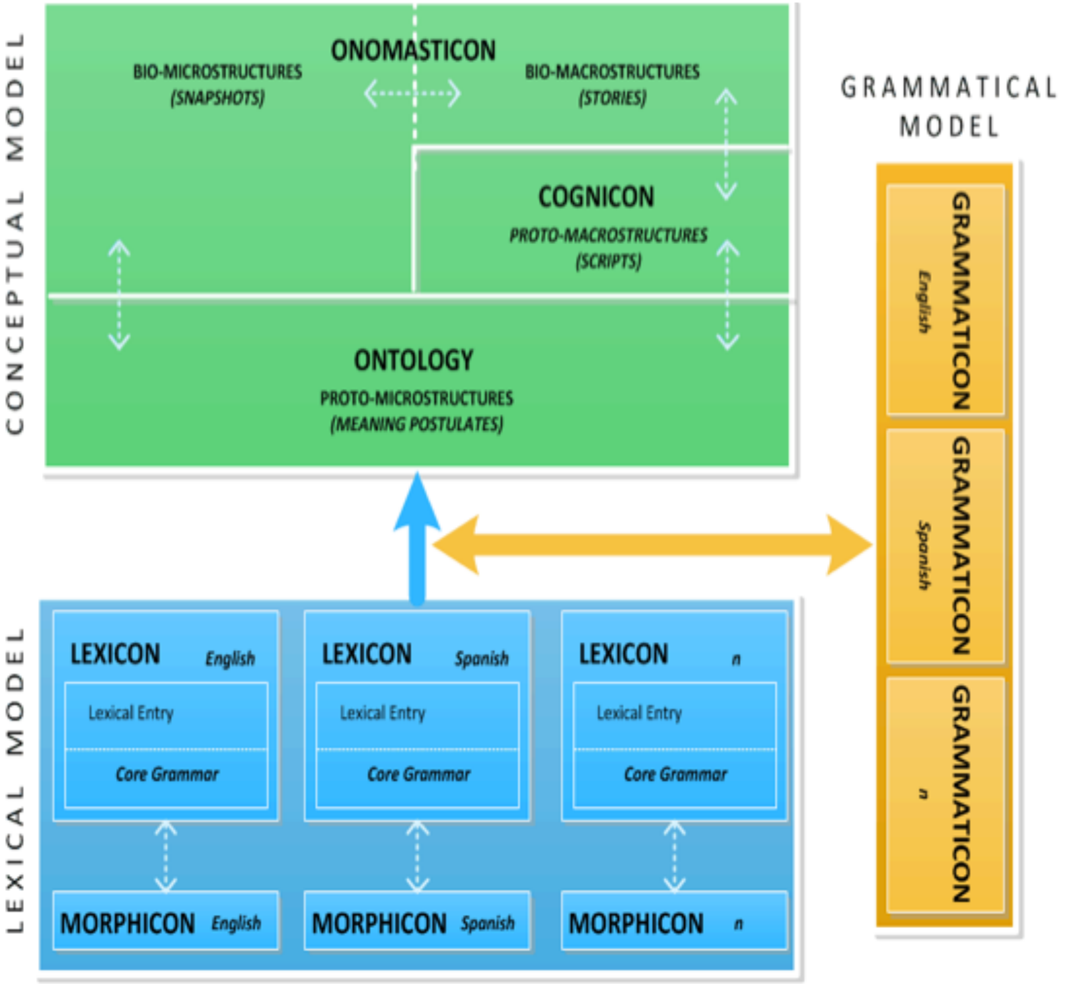
> e_1: cocaine is a drug; e_2/e_3: it is manufactured with coca leaf; e_4: it gives you pleasure and energy; e_5: it is generally insufflated; e_6: (in many countries) it is illegal to consume or traffick with it.

The *genus*, the concept +DRUG_00, will always be included in the meaning postulate, in the first predication in particular (e_1). The meaning postulates consist of one or more generic predications (e_1, e_2, ... e_n), whose function is that of describing the commonplaces which make up our knowledge of the world. One could say that a predication is applied to "all the typical entities", that is to say, those entities which have charac-

[7] We refer the reader to Periñán & Mairal (2010) for examples of conceptual representation in the form of thematic frames and meaning postulates.

[8] This example comes from the *Globalcrimeterm* subontology under construction and was presented by the author in the 2013 *Role and Reference Grammar* International Conference in Freiburg, Germany.

FIGURE 1
FunGramKB architecture



teristic features. People observe a series of regularities in the world around us which they use to predict the actions of other people or changes in our environment. In short, the language of representation, COREL, allows us to give definitions of each ontological concept. But how can we retrieve the extralinguistic information referred to above from these definitions? All these definitions serve as inductions to a reasoning engine which allows the computer to simulate human reasoning patterns and thus come to conclusions using the same unspecialised knowledge about things from everyday life. The working of this internal reasoning engine is what we call *MicroKnowing* (*Microconceptual Knowledge Spreading*).

3.3. Legal ontologies and subontologies

The conceptual apparatus presented above would be beneficial to the potential development, within FunGramKB, of specialised subontologies which interact with the general cognitive model (the Core Ontology as the main actor, the onomasticon and the cognicon) and the lexical model, possibly in each of the integrated lexicons (Spanish, English and Italian). The subontology chosen for this research project lies within the framework of the legal field and criminal law, even if many concepts are shared with the general knowledge of the world of non-expert speakers. Precedents to ontologies in general are very wide (see e.g. Musen, 1992, or Gruber, 1993), although we should go back first to the onomasiological lexicography works collected by Martín-Mingorance (1994), which, although coming from philosophy and bearing no relation with the concept of ontology in NLP, they do allow for the human endeavour to comprehend the structure of knowledge and reality[9]. Finally, FunGramKB also benefits from other developments based on onomasiological lexicology which deal with the axiological parameter integration in the knowledge base and the axiological aspects of criminal law (Felices, 2010). All these currents of thought are perfectly integrated into FunGramKB, which has increased its scope from all the proposals referred to above over the years.

Concerning the precedents of ontologies used in the legal field one should cite Valente (2005); Breuker, Valente & Winkels (2005); Breuker, Casanovas, Klein & Francesconi (2008), or Sartor, Casanovas, Biasiotti & Fernández-Barrera (2011), which can be seen below. Nonetheless, all authors in general point out the problem of defining as ontologies elaborations which are very different to one another and also the difficulty to distinguish between ontologies, in the strict sense of the word, knowledge representations or knowledge bases, although at times different elements may be combined. In addition, Periñán & Arcas (2007) assert that the large majority of these "misnamed" ontologies are, in fact, lexical taxonomies which do not give formal representation of meaning to each of their terms, but which are rather infradefined as regards their subsumptive relation with other terms (and sometimes with other semantic relations such as synonymy, meronymy, etc.). Some of the so called ontologies:

(i) organize and structure information, as in the case of projects such as *Jur-Wordnet* (Gangemi, Sagre & Tiscornia, 2005) or the Italian ontology of crimes (Asaro et al. 2003; Lenci, 2008);

(ii) have a reasoning and a problem solving engine, such as the ontology *CLIME* for maritime law (Boer, Hoekstra & Winkels, 2001) or *Argument Developer*, which works with different types of legal data bases (Zeleznikow & Stranieri, 2001);

(iii) have semantic indexing and search, such as the ontologies of French codes (Lame, 2002), ontologies which represent cases of financial fraud (Leary, Vandenberghe & Zeleznikow, 2004) or which develop an intelligent FAQ (Frequently Asked Questions) system for judges (Benjamins et al., 2003; Casanovas, Casellas & Vallbé, 2008);

(iv) understand a domain, such as those which are more generally applied in law, e.g. the functional ontologies of law (based on *Ontolingua*) by Valente and Breuker (1994, 1995, 1999), and those of language of legal discourse by McCarty (1989) or those more general ontologies used for knowledge representation (*Frame Ontology*) by Van Kralingen (1995). They all use general language for expressing legal knowledge.

The number of researchers who are working at present on legal ontologies is very high, although as far as we have observed the numerous applications which have existed up to now

[9] This author refers to works by universal figures such as Aristotle and his philosophy of essence, *the Porphyrian Tree*, Pliny the Elder's *Natural History*, Isidore of Seville's *Etymologiae*, or, in the early modern period, Francis Bacon's *Instauratio Magna* and his *Novum Organum*, or Comenius's *Ianua Linguarum Reserata*. In the contemporary period one should highlight *Roget's Thesaurus* and the work of the Scotsman, Wilkins. These are just some of the precedents to more modern and useful ideological dictionaries, such as the *Longman Lexicon of Contemporary English* by McArthur, and in Spanish, the well-known *Diccionario ideológico* by Julio Casares.

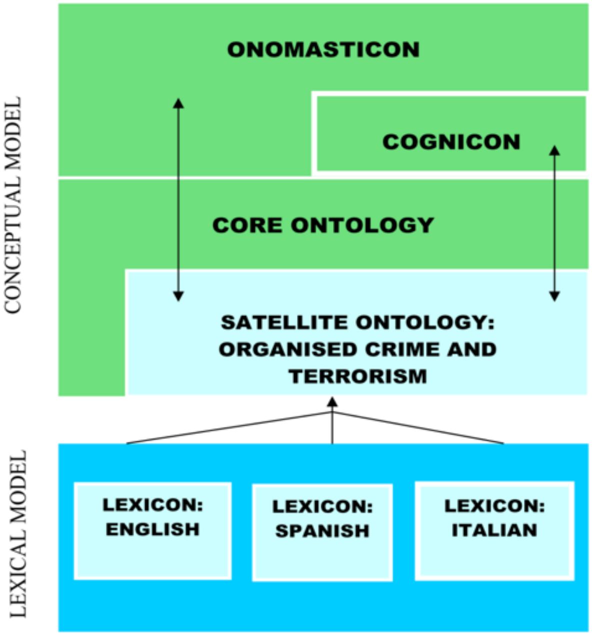
are based on models which lie far from the architecture offered by FunGramKB and do not contain any subontological development which cover terrorism and organized crime stemming from a knowledge base connected to a constructional linguistic model.

4. Major assumptions and procedural details

This subontology can be defined as a hierarchical taxonomy of specialised concepts belonging to an expert area of knowledge: basically, an area of criminal law. It thus serves the purpose of enhancing FunGramKB with specialised knowledge, as the knowledge base has been so far implemented to work with elementary common-sense concepts of human cognition. The Core Ontology and the Satellite Ontologies become connected as shown in figure 2:

FIGURE 2

Extension of FunGramKB architecture including Satellite Ontologies



The resulting repository on the topic selected dealing with criminal law might benefit both humans (by means of an interface acting as a dictionary) and machines, since it can be applied to Natural Language Processing systems (NLP) in future stages of the programme. For this purpose, it has already been noted that a semi-automatic population of the ontology and its hierarchical structure is essential, involving the building of relevant meaning postulates and the creation of terminal concepts (and subconcepts); this process must be followed by the semi-automatic population of specialised lexica for the languages chosen, which cover relevant lexical information and which can store hundreds of lexical units into each lexicon.

In order to feed the Satellite Ontology and the Lexica, terminological units must be obtained from documentary and textual databases offered from reliable reference sources, such as regulations, treaties, articles, books, glossaries or previous legal ontologies provided by international agencies and institutions which work on criminal law or the fight against organized crime such as EUROPOL, EUROJUST or CDPC (*European Committee on Crime Problems*), among others. The overall guidelines to process the information will be explained below (section 4.2), but the design of *FunGramKB Term Extractor* (Periñán & Arcas, 2014; Felices & Ureña, 2014) has been the pivotal element integrated in *FunGramKB Suite* to deal with the *Globalcrimeterm* corpus (Felices & Ureña, 2012). This is the basic instrument which allows not only the automatic identification of candidate terms according to their probabilistic weight, but also the technical support to terminologists to choose the relevant terms for the Satellite Ontology.

4.1. Specific objectives

As previously stated in section 2, the final objective of this project is to design a subontological model with a conceptual base and to implement it in the criminal law module in the domain of terrorism and organised crime, as well as to populate its corresponding (English, Spanish and Italian) terminological lexica within FunGramKB's architecture. In order for this to happen, the following specific objectives must be fulfilled:

1) The organisation of the core conceptual structure of criminal law must be carried out by analysing the evidence obtained in the epistemological frameworks of previous legal ontologies described in section 3.3 and also by seeking expert advice through professionals and by consulting specialised sources. These structures, based on the deductive approach to the thematic domain will be verified or substituted through the inductive methodological phase of the project. Thus, the field work will either ratify or not ratify the coherence and reliability of the structure initially proposed. This process is known in the practice of the elaboration of legal ontologies as 'common sense'.
2) The defining words of the thematic domain in the field of criminal law (organised crime and terrorism) must be identified through intensive searches in the most relevant digital sources and resources at our disposal. In this way defining texts can be processed more easily and automatically. The extractor (FTE) is the specific tool to facilitate and conduct this process.
3) The information must be included in FunGramKB through its online editor, which will connect, on the conceptual level, the subontology which we will have created with the Core Ontology, the cognicon and the onomasticon; and, on the lexical level, the lexica corresponding to the three languages selected for this project.
4) A specific dictionary-like interface for this terminological subontology must be designed and its information must be able to be accessed both by humans and by the machine through the language of conceptual representation COREL.
5) The ontology must be semi-automatically populated in its hierarchical structuring: appropriate meaning postulates must be constructed and domain-related basic concepts, terminal concepts (and subconcepts) must be created.
6) Specialised lexica corresponding to the English, Italian and Spanish languages must be semi-automatically populated with their pertinent lexical information at a minimum rate of 800 lexical units per lexicon.
7) The final product will be obtained after the preceding populations: a repository will have been created on knowledge about criminal law concerning terrorism and organised crime for its potential exploitation in tasks of automatic translation and retrieval of information for organisms which deal with international cooperation in the above-mentioned topics.

4.2. Relevant tasks

With the aim of fulfilling the objectives presented in the previous section, the following methodology has been planned:

1. ***Task:*** The highest number of terminological resources must be looked for (e.g. monolingual dictionaries, thesauri, lexical taxonomies, etc.) in English, Italian and Spanish on criminal law, and the fields of terrorism and organised crime. Digital resources will be preferred as this will enable us to process content automatically, i.e. tasks of tokenization and lemmatization, etc. ***Method:*** In this activity, research teams will independently and contrastively filter the compilation of all possible available paper and digital sources concerning the English, the Italian and the Spanish language, regarding their possible usefulness for the project's objectives and following a list of previously established criteria. At this point consultations will also be made to external experts in order to be as ex-

haustive as possible with the included and selected resources. The *globalcrimeterm* corpus is the result of this process[10].

2. ***Task:*** An inventory of basic defining vocabulary will be elaborated from the *definiens* found in the previous lexicographical resources. The main criteria will be to make a frequency index of the words which constitute the defining texts once their functional words (i.e. articles, prepositions, etc.) have been removed. To facilitate this process the *FunGramKB Term Extractor* is the necessary instrument. ***Method:*** In this task the main participants will be jurists and selected linguists. Other members of the research group may contribute with suggestions and proposals, especially those specialists in English-Spanish, English-Italian and Spanish-Italian legal translation and in the application of functional grammar models to specialised terminology. For this task similar templates to those used for the same purpose in the elaboration of the FunGramKB's Core Ontology will be used. It must be borne in mind here that this step is closely connected with the next one and requires a clear discernment between what is involved at a linguistic level of knowledge (lexical units) and at a conceptual one (concepts).

3. ***Task:*** A multilingual hierarchical arrangement of the defining terms will be carried out according to the taxonomical subsumptive relation (IS-A). In order to do this, the prior conceptualisation of the terminiological inventory will be necessary, i.e. the projection of the terms onto conceptual units. In this sense, different phenomena, characteristic of the conceptual model, will be produced: terms which are grouped under the same concept, concepts which present lexical gaps in certain languages, etc. ***Method:*** In short, the same methodology employed in the structuring of the basic conceptual level of the Core Ontology will be used: i.e. conceptualisation, hierarchization, remodelling and refining. In other words, the COHERENT methodology (Periñán & Mairal, 2011).

4. ***Task:*** A meaning postulate must be specified for each and every one of the basic concepts of the terminological subontology. In this way, we can also check the validity of the subsumptive relations established through the genus concept of the meaning postulates. Special emphasis will be given to the detail of the information presented in these definitions and in the preciseness of their formal representation with COREL (*Conceptual Representation Language*). ***Method:*** One must bear in mind that the system of analysis will follow a procedure inspired by Simon C. Dik's stepwise lexical decomposition (1989) which we will call "stepwise conceptual decomposition", although instead of referring to lexical units it will be applied to concepts. This *modus operandi* could be defined as a process through which conceptual units of a predicate are replaced with their respective meaning postulates until a representation of the meaning is reached which is made up of metaconceptual primitives. The implications between the lexical and grammatical model on the one hand and the conceptual one on the other at this point require the simultaneous cooperation of the specialists in NLP and the linguists.

5. ***Task:*** The root concepts in the basic conceptual level of the terminological subontology will be connected with a concept (either basic or terminal) of FunGramKB's Core Ontology in order to minimise informative redundancy and maximise the capacity of information and knowledge transmission.

[10] This domain-specifc corpus comprises 621 documents and 5.698.754 tokens (Felices & Ureña, 2012; Periñán & Arcas, 2014).

Method: One must bear in mind that the option of using the methodology of deep semantics in order to develop knowledge bases requires the building of a language of representation (COREL) which allows us to define conceptual units, with the additional benefit that the conceptual relations can be equally obtained by applying mechanisms of inheritance and inference on the meaning postulates. Eventually, this process will facilitate the phase of connection between the Core Ontology and the subontology and requires a close cooperation between the NLP experts and the linguists.

6. ***Task:*** The terminological subontology will be populated with terminal concepts, also specifying their meaning postulates. ***Method:*** This phase, together with the next, are the longest and require the greatest time commitment on the part of the researchers, according to the previously defined conceptual areas in preceding tasks and which correspond to the conceptual hierarchy in the criminal law framework regarding terrorism and organised crime. The populating of terminal concepts will follow a similar procedure to that followed previously for the populating of terminal concepts in FunGramKB for the Core Ontology. Basically, once the ability to edit concepts is empowered in the editor, the steps to follow will be the following: (1) to open all the online resources with the resources obtained before in the multilingual context (English, Italian and Spanish); (2) to access the conceptual domain which is to be populated, and (3) to choose at least one of the basic concepts and search for possible sources of this concept in the different corpora, paying attention to the possible differentiating parameters which will help to form the terminal concepts. Given the magnitude of the possible concepts, this task should be shared by all the specialists involved in the project.
7. ***Task:*** The lexical entries of the terms assigned to the new concepts of the terminological subontology will be built up. ***Method:*** For this stage specialised corpora will be used to identify those constructions which could intervene in the headword terms of each lexical entry. In this case the procedure to follow is similar to that of the previous stage, but given that we are working with a lexical module of the LCM it will be necessary to complete the lexical templates which correspond to each entry in each of the three target languages of this specific project.
8. ***Task:*** FunGramKB's reasoning engine will be applied to the resulting subontology in order to check that the results expected have been obtained as regards information inheritance and inference. The specialists in Computational Linguistics will be in charge of this last stage which completes the full activation of the terminological subontology.

5. Conclusions

Terminological subontologies compatible with FunGramKB must be developed in three phases as described in the objectives and tasks above (sections 4.1 and 4.2). The first stage involves the compilation of a collection of specialized texts, which are used as a corpus for the identification of terminology, that is, the linguistic units which carry specialised knowledge related to the field towards which conceptualisation is targeted. The second stage consists of the extraction of terminology from the corpus and, finally, the third stage includes conceptual modelling tasks. One of the main claims of the research project presented in this article has been precisely to contribute with a stepwise methodology for the construction of subontologies, which is applicable in the modelling of any specialized domain of knowledge, regardless of the corpus or corpora that are being used as the linguistic input. The construction of domain-specific ontologies in this vein will allow FunGramKB to direct these developments to future tasks related to data mining, information retrieval, or the resolution of problems in which intensive reasoning is involved.

In my opinion, the key role that the building of a legal subontology like *Globalcrimeterm* may play has been established in theoretical terms in this article. In practical terms, however, the conceptual modelling and hierarchization phases of this research are proving that the area of international cooperation in criminal law is still an admittedly small area of expert knowledge and, paradoxically, the results obtained so far[11] are confirming, on the one hand, the multidisciplinary nature of this field and, on the other, the evidence that too many lexical units or multiword expressions in this domain are widely used by the general public or excessively shared with the experts. Therefore, more work is required in order to conceptualise broader areas of the legal field. In the same vein, there are still formidable challenges to be faced and new research will be necessary for a comprehensive development of new specialised ontologies in FunGramKB, particularly from other scientific disciplines, such as life, natural or formal sciences.

6. Bibliographic references

Aguado de Cea, Guadalupe, Elena Montiel & José Á. Ramos, 2007: "Multilingualidad en una aplicación basada en el conocimiento", *Procesamiento del Lenguaje Natural* 38, 77-98.

Allen, James F., 1983: "Maintaining knowledge about temporal intervals", *Communications of the ACM* 26 (11), 832-843.

Allen, James F. & George Ferguson, 1994: "Actions and events in interval temporal logic", *Journal of Logic and Computation* 4 (5), 531-579.

Asaro, Carmelo et al., 2003: "A Domain Ontology: Italian Crime Ontology", *Proceedings of the ICAIL 2003 Workshop on Legal Ontologies & Web Based Legal Information Management*, 1-7.

Benjamins, V. Richard et al., 2003: "Ontologies of Professional Legal Knowledge as the Basis for Intelligent IT Support for Judges", *Proceeedings of the ICAIL 2003 Workshop on Legal Ontologies & Web based legal information management*.

Boer, Alexander, Rinke Hoekstra & Radboud Winkels, 2001: "The CLIME Ontology", *Proceedings of the Second International Workshop on Legal Ontolologies*, Amsterdam: Jurix, 37-47.

Breuker, Joost, André Valente & Radboud Winkels, 2005: "Use and reuse of legal ontologies in knowledge engineering and information management" in Richard Benjamins, Pompeu Casonovas, Joost Breuker & Aldo Gangemi (eds.): *Law and the semantic web*, vol. 3369, Berlin: Springer, 36-64.

Breuker, Joost, Pompeu Casanovas, Michel C. A. Klein & Enrico Francesconi (eds.), 2008: *Law, ontologies and the semantic web. Channelling the legal information flood, Frontiers in Artificial Intelligence and Applications*, vol. 188, Amsterdam: IOS Press.

Casanovas, Pompeu, Nuria Casellas & Joan-Josep Vallbé, 2008: "An ontology-based Decision Support System for Judges" in Joost Breuker, Pompeu Casanovas, Michel C. A. Klein & Enrico Francesconi (eds.): *Legal Ontologies and the Semantic Web. Channeling the Legal Information Flood*, Amsterdam: IOS Press, 165-175.

Dik, Simon C., 1989: *The Theory of Functional Grammar. Part I: The Structure of the Clause*, Dordrecht: Foris.

[11] For this purpose, the reader can see the contributions of this project published so far by clicking on the link *Research*, then on *Publications* in: www.fungramkb.com.

Felices, Ángel, 2010: "Axiological Analysis of Entries in a Spanish Law Dictionary and Their English Equivalents", *Researching Language and the Law: Textual Features and Translation Issues*, Bern: Peter Lang.

Felices, Ángel & Pedro Ureña, 2012: "Fundamentos metodológicos de la creación subontológica en FunGramKB", *Onomázein* 26, 49-67.

Felices, Ángel & Pedro Ureña, 2014: "FunGramKB Term Extractor: a key instrument for building a satellite ontology based on a specialized corpus" in Brian Nolan & Carlos Periñán (eds.): *Language processing and grammars: The role of functionally oriented computational models* (Studies in Language Series), Amsterdam: John Benjamins, 251-269.

Gangemi, Aldo, Maria-Teresa Sagri & Daniela Tiscornia, 2005: "A Constructive Framework for Legal Ontologies" in Richard Benjamins et al. (eds.): *Law and the Semantic Web*, Berlin: Springer, 97-124.

Gruber, Thomas R., 1993: "A translation approach to portable ontologies", *Knowledge Acquisition* 5(2), 199-220.

Lame, Guiraude, 2002: *Construction d'ontologie à partir de textes. Une ontologie de droit dédiée à la recherche d'information sur le Web*. PhD dissertation, Ecole des mines de Paris, Paris [http://www.cri.ensmp.fr/].

Leary, Richard, Wim Vandenberghe & John Zeleznikow, 2004: "Towards a financial fraud ontology: a legal modelling approach", *ICAIL 2003 Workshop on Legal Ontologies & Web based legal information management*, 1-33 (to unload it, go to Leibnizcenter.org).

Lenci, Alessandro (ed.), 2008: *From context to meaning: distributional models of the lexicon in linguistics and cognitive science, Italian Journal of Linguistics* 20/1 (special issue).

Mairal, Ricardo & Francisco J. Ruiz de Mendoza, 2008: "New challenges for lexical representation within the Lexical-Constructional Model", *Revista Canaria de Estudios Ingleses* 57, 137-158.

Mairal, Ricardo & Francisco J. Ruiz de Mendoza, 2009: "Levels of description and explanation in meaning construction" in Christopher Butler & Javier Martín (eds.): *Deconstructing constructions*, Amsterdam: John Benjamins, 153-198.

Mairal, Ricardo & Carlos Periñán, 2009: "The anatomy of the lexicon component within the framework of a conceptual knowledge base", *Revista Española de Lingüística Aplicada* 22, 217-244.

Mairal, Ricardo & Carlos Periñán,2010: "Role and Reference Grammar and Ontological Engineering" in José Luis Cifuentes et al. (eds.): *Los caminos de la lengua: Estudios en homenaje a Enrique Alcaraz Varó*, Alicante: Universidad de Alicante, 649-65.

Martín-Mingorance, Leocadio, 1994: "La lexicografía onomasiológica" in Humberto Hernández, *Aspectos de lexicografía contemporánea*, Barcelona: Biglograf.

Martín-Mingorance, Leocadio, 1998: *El modelo lexemático-funcional* [posthomous work edited by Amalia MARÍN], Granada: Universidad de Granada.

McCarty, L. Thorne, 1989: "A Language for Legal Discourse, I. Basic Features", *Proceedings of the 2nd International Conference on Artificial Intelligence and Law*, New York: ACM, 180-189.

Mcguinness, Deborah L. et al., 2000: "An Environment for Merging and Testing Large Ontologies", *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning* (KR2000), Breckenridge, Colorado, [http://disi.unitn.it/~accord/RelatedWork/Matching/McGuinnessKR.pdf, consultation date: February 25, 2015].

Montiel, Elena et al., 2007: "Localizing ontologies in OWL", *Workshop at the 6th International Web*

Conference. From Text to knowledge, Busan, Korea, 13-22.

Musen, Mark A., 1992: "Dimensions of knowledge sharing and reuse", *Computers and Biomedical Research* 25, 435-467.

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

Periñán, Carlos & Francisco Arcas, 2005: "Microconceptual-Knowledge Spreading in FunGramKB", *Proceedings on the 9th IASTED International Conference on Artificial Intelligence and Soft Computing*, Anaheim-Calgary-Zurich: ACTA Press, 239-244.

Periñán, Carlos & Francisco Arcas, 2006: "Reusing computer-oriented lexica as foreign-language electronic dictionaries", *Anglogermánica Online* 4, 69-93.

Periñán, Carlos & Francisco Arcas, 2007: "Cognitive modules of an NLP knowledge base for language understanding", *Procesamiento del Lenguaje Natural* 39, 197-204.

Periñán, Carlos & Francisco Arcas, 2008: "A cognitive approach to qualities for NLP", *Procesamiento del Lenguaje Natural* 41, 137-144.

Periñán, Carlos & Francisco Arcas, 2010: "Ontological commitments in FunGramKB", *Procesamiento del Lenguaje Natural* 44, 27-34.

Periñán, Carlos & Francisco Arcas, 2014: "La ingeniería del conocimiento en el dominio legal: La construcción de una Ontología Satélite en FunGramKB", *Revista Signos. Estudios de Lingüística* 47(84), 113-139.

Periñán, Carlos & Ricardo Mairal, 2009: "Bringing Role and Reference Grammar to natural language understanding", *Procesamiento del Lenguaje Natural* 43, 265-273.

Periñán, Carlos & Ricardo Mairal, 2010: "Enhancing UniArab with FunGramKB", *Procesamiento del Lenguaje Natural* 44, 19-26.

Periñán, Carlos & Ricardo Mairal, 2011: "The COHERENT methodology in FunGramKB", *Onomázein* 24, 13-33.

Ruiz de Mendoza, Francisco J. & Ricardo Mairal, 2008: "Levels of description and constraining factors in meaning construction. An introduction to the Lexical Constructional Model", *Folia Linguistica* 42 (2), 355-400.

Ruiz de Mendoza, Francisco J. & Ricardo Mairal, 2011: "Constraints on syntactic alternation: lexical-constructional subsumption in the Lexical Constructional Model" in Pilar Guerrero (ed.): *Morphosyntactic Alternations in English: Functional and Cognitive Perspectives*, London: Equinox.

Sartor, Giovanni, Pompeu Casanovas, Maria Angela Biasotti & Maritxel Fernández-Barrera (eds.), 2011: *Approaches to legal ontologies, theories, domains, methodologies*, Berlin: Springer.

Valente, 2005: "Types and roles of legal ontologies", in Richard Benjamins, Pompeu Casanovas, Joost Breuker & Aldo Gangemi (eds.): *Law and the semantic web*, vol. 3369, Berlin: Springer, 65-76.

Valente, Andre & Joost Breuker, 1994: "A functional view of law" in Gabriella Bargellini & Simona Binazzi, (eds.): *Towards a global expert system in law*, Padua: CEDAM Publishers, 12-24.

Valente, Andre & Joost Breuker, 1999: "Legal Modelling and Automated Reasoning with ONLINE", *International Journal of Human-Computer Studies* 51(6), 1079-1125.

Van Kralingen, Robert W., 1995: *Frame-based conceptual models of statute law*, The Hague: Kluwer Law International.

Van Valin, Robert D. Jr., 2005: *Exploring the syntax-semantics interface*, Cambridge: Cambridge University Press.

Van Valin, Robert D. Jr., & Randy J. LaPolla, 1997: *Syntax, structure, meaning and function*, Cambridge: Cambridge University Press.

Velardi, Paola et al., 1991: "How to encode semantic knowledge: a method for meaning representation and computer-aided acquisition", *Computational Linguistics* 17/2, 153-170.

Zeleznikow, John & Andrew Stranieri, A. 2001: "A framework for the construction of legal decision support systems", *Proceedings of the Fifth Business Information Systems Conference*, Calgary, Canada: ACM Press, 240-250.