RESLA 22 (2009), 217-244

THE ANATOMY OF THE LEXICON WITHIN THE FRAMEWORK OF AN NLP KNOWLEDGE BASE

RICARDO MAIRAL USÓN*
UNED Madrid
CARLOS PERIÑÁN-PASCUAL
Universidad Católica San Antonio, Murcia

ABSTRACT. *The aim of this paper is to present the format of the lexical level within the framework of FunGramKB (www.fungramkb.com), a lexical conceptual knowledge base that is part of the Lexical Constructional Model (www.lexicom.es). In doing so, we discuss the different features that define the Spanish and the English lexica*.
KEY WORDS. *Ontology, lexicon, conceptual logical structures,constructions, aktionsart.*

RESUMEN. *El objetivo de este trabajo es presentar el formato del nivel léxico en el contexto de la base de conocimiento léxico conceptual FunGramKB (www.fungramkb.com) que, a su vez, forma parte del Modelo Léxico Construccional (www.lexicom.es). Ofrecemos una descripción de los rasgos esenciales que definen el componente léxico en español e inglés.*
PALABRAS CLAVE. *Ontología, léxicón, estructuras lógico conceptuales, construcciones, aktionsart.*

1. INTRODUCTION

The Lexicom research group[1] has developed the Lexical Constructional Model (LCM), a usage-based comprehensive theory of meaning construction that aims to explain how all aspects of meaning construction including those that go beyond core grammar (e.g. traditional implicature, illocutionary force, and discourse coherence) interact among one another (cf. Mairal and Ruiz de Mendoza, 2008; Ruiz de Mendoza and Mairal, 2008). Hence, if the output of the LCM is a fine-grained delicate description of all aspects involved in meaning construction, then the LCM offers a very nice framework for the development of natural language processing (NLP) applications based on a deep semantic approach. In connection with this, we have expanded and adapted

previous work by Periñán-Pascual and Arcas-Túnez (2004, 2005, 2006, 2007a, 2008a, 2008b) and have developed an updated version of FunGramKB, a lexico-conceptual knowledge base that integrates very rich semantic and syntactic information that allows the development of natural language applications.

Within this context, this paper addresses the format of the lexicon component as conceived in FunGramKB. Recall that the construction of computational lexica can be carried out by means of two methods: creation or acquisition (Calzolari and Picchi 1994). The former typically involves the construction of lexica out of lexicographers' introspection, while the latter occurs in an automatic or assisted way with the aid of electronic language resources, e.g. dictionary or corpora. Velardi *et alii* (1991) notice that the hand-made construction of computational lexica presents some drawbacks, such as heavy workload and lack of both consistency and sistematicity, which turn out to be determining factors as the size of the lexicon increases. Another alternative consists in reusing existing language resources with the aim of acquiring lexical knowledge in a (semi-)automatic way. In this respect, machine-readable dictionaries (MRD) become one of the most useful resources. The next issue is to decide whether the process of information extraction should be automatic (i.e. through an algorithm) or interactive (i.e. through a lexicographic tool).

One of the main limitations of automatically exploiting the potential of MRDs lies in the nature itself of lexicography. Firstly, dictionaries are primarily designed for humans, so lexicographers rely on readers' linguistic competence in order to minimize the amount of data in lexical entries. The problem is that dictionaries usually ignore basic facts of commonsense knowledge, which play a key role in NLP text understanding. Secondly, lexicographers work under pressures of time and space, which favour the inconsistency of entries. For example, lexical units with a similar morphological, syntactic and/or semantic behaviour are not treated in a similar way. Consequently, determining metatextual variants in MRDs actually results in more workload than the hand-made construction of computational lexica (Ide and Véronis 1994a). Thirdly, polysemy is not handled adequately, since the various senses of lexical units are simply enumerated. Although some research performed with MRDs has been successful, the amount of semantic information which can be automatically extracted from lexicographic definitions does not fully meet the needs of NLP, producing thus little more than a handful of limited and imperfect taxonomies (Ide and Véronis 1994b).

Nowadays the accessibility to a wide range of electronic language resources can make fully-automatic lexical acquisition be an appealing strategy. However, we conclude that the time involved for this task can be similar to the time taken for the computer-assisted construction of computational lexica. In this respect, FunGramKB Lexicon Editor displays a user-friendly interface which allows knowledge engineers to develop large-scale lexica consistently.

Then, the organization of this paper has the following format. Section 2 provides a selective description of the knowledge base with a special emphasis on the elements and the properties of the ontology. A first approximation to the ontology is essential to

understand the actual format of the lexicon, since lexical entries are heavily influenced by the ontology. Section 3 describes the features in FunGramKB lexical entries. Finally, section 4 presents a few concluding remarks[2].

2. FUNGRAMKB

FunGramKB Suite[3] is a user-friendly online environment for the semiautomatic construction of a multipurpose lexico-conceptual knowledge base for NLP systems. On the one hand, FunGramKB is multipurpose in the sense that it is both multifunctional[4] and multilingual[5]. In other words, FunGramKB has been designed to be reused in various NLP tasks (e.g. information retrieval and extraction, machine translation, dialogue-based systems, etc) and with several natural languages.[6] On the other hand, our knowledge base comprises two general levels of information: a lexical level and a conceptual level. What follows is a description of FunGramKB architecture and an account of the main ontological elements and properties in this knowledge base.

2.1. *FunGramKB architecture*

As stated above, FunGramKB is made up of two information levels, which in turn consist of several independent but interrelated modules:

Lexical level (i.e. linguistic knowledge):

- The lexicon stores morphosyntactic, pragmatic and collocational information about lexical units.[7] FunGramKB and the LCM share their lexical model. However, this model is not a literal implementation of the lexical database in Role and Reference Grammar (RRG), since some important contributions have been introduced with the aim of building a robust knowledge base, although the major linguistic assumptions of RRG are preserved, i.e. logical structures, macroroles, and the rest of the linking algorithm.
- The morphicon helps our system to handle cases of inflectional morphology.

Conceptual level (i.e. non-linguistic knowledge):

- The ontology is presented as a hierarchical catalogue of all the concepts that a person has in mind when talking about everyday situations. Here is where semantic knowledge is stored in the form of meaning postulates.
- The cognicon stores procedural knowledge (e.g. how to fry an egg, how to buy a product, etc.) by means of cognitive macrostructures, i.e. script-like 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 and Ferguson, 1994).

- The onomasticon stores information about instances of entities and events, such as Bill Gates, Taj Mahal, or 9/11. This module stores two different types of schemata (i.e. snapshots and stories), since instances can be portrayed synchronically or diachronically.

The main consequence of this two-level design is that every lexical module is language-dependent, while every conceptual module is shared by all languages involved in the knowledge base. Therefore, computational lexicographers must develop one lexicon and one morphicon for English, one lexicon and one morphicon for Spanish and so on, but knowledge engineers build just one ontology, one cognicon and one onomasticon to process any language input cognitively. This paper then focuses on the actual format of the different lexica involved.

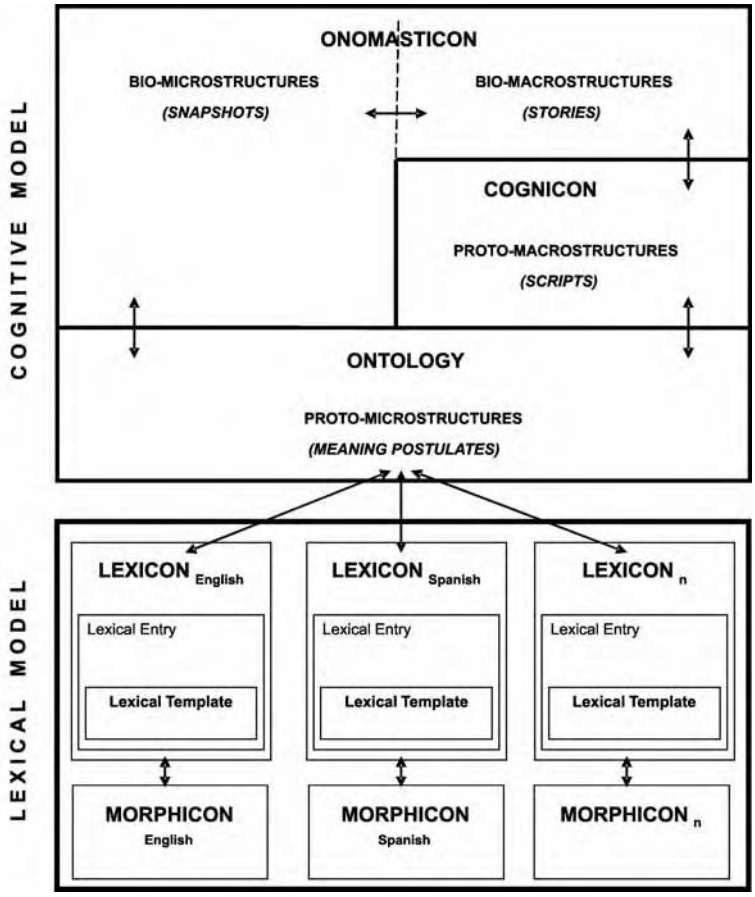

Figure 1. *FunGramKB modules.*

According to Figure 1, FunGramKB lexica are populated in a top-down fashion, i.e. the construction of lexical entries involves the previous ontological modelling of their corresponding concepts. Thus, providing that some knowledge engineer has introduced the concept +DRY_01 together with its thematic frame and meaning postulate (cf. sections 2.2.2. and 2.2.3), then linguists are able to type in information concerning the lexical units *dry* or *secar*.

As shown here, the ontology becomes the pivot for the different lexica, which explains why we maintain that this model is conceptually rather than lexically-driven. More importantly, this type of approach provides a nice framework to deal with one of the most controversial issues in lexical representation, i.e. the lexico-conceptual linkage (cf. Mairal and Periñán, 2009). Next section describes the key module in FunGramKB architecture, i.e. the ontology.

2.2. *FunGramKB ontology*

FunGramKB is provided with a universal, linguistically-motivated and general-purpose ontology. Firstly, our ontology takes the form of a universal concept taxonomy, where "universal" means that every concept[8] we imagine has, or can have, an appropriate place in this ontology (Corcho, Fernández López and Gómez Pérez 2001). This term can also be applied to FunGramKB in the sense that we adopt a universal approach on the relation between language and conceptualization, where interlingual differences in syntactic constructions do not involve conceptual differences (cf. Jackendoff 1983, 1990). In our case, the relation between language structures and conceptual constructs is mediated by what we have called *conceptual logical structures*, where phenomena such as diathetic alternations are directly reflected (cf. sections 3.1. and 4).

Secondly, FunGramKB ontology is linguistically motivated, as a result of its involvement with the semantics of lexical units, although the knowledge stored in our ontology is not specific to any particular language. This is the reason why a new concept should be introduced in our ontology whenever there is at least one lexical unit whose meaning does not match any of the meaning postulates stored in the knowledge base provided that the values of the ontological properties of that new concept are shared by all lexical units which are linked to it.

Finally, our ontology is general-purpose, because neither it is domain-specific nor contains terminological knowledge[9].

Nowadays there is no single right methodology for ontology development. Ontology design tends to be a creative process, so it is probable that two ontologies designed by different people have a different structuring (Noy and McGuinness, 2001). Thus, the ontology model should be founded on a solid methodology, which contributes to avoid some common errors in conceptual modelling. Although methodological criteria applied to FunGramKB ontology are presented elsewhere (cf. Periñán-Pascual and Arcas-Túnez 2007a), in the remaining of this section we describe the main features of the conceptual elements and the semantic properties in our ontology.

2.2.1. Ontological elements

FunGramKB ontology distinguishes three different conceptual levels, each one of them with concepts of a different type: metaconcepts, basic concepts and terminals.

Metaconcepts, preceded by symbol # (e.g. #ABSTRACT, #COLLECTION, #EMOTION, #POSSESSION, #TEMPORAL etc), constitute the upper level in the taxonomy. The analysis of the upper level in the main linguistic ontologies—DOLCE (Gangemi et al., 2002; Masolo et al., 2003), Generalized Upper Model (Bateman, 1990; Bateman, Henschel and Rinaldi, 1995), Mikrokosmos (Beale, Nirenburg and Mahesh, 1995; Mahesh and Nirenburg, 1995; Nirenburg et al., 1996), SIMPLE (Lenci, 2000; Lenci et al., 2000; Pedersen and Keson, 1999; SIMPLE Specification Group, 2000), SUMO (Niles and Pease, 2001a, 2001b)—led to a metaconceptual model whose design contributes to the integration and exchange of information with other ontologies, providing thus standardization and uniformity. The result amounts to forty-two metaconcepts distributed in three subontologies: #ENTITY, #EVENT and #QUALITY.[10]

Basic concepts, preceded by symbol + (e.g. +BOOK_00, +DIRTY_00, +FORGET_00, +HAND_00, +MOVE_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. The starting point for the identification of basic concepts was the defining vocabulary in *Longman Dictionary of Contemporary English* (Procter, 1978) and *Diccionario para la Enseñanza de la Lengua Española* (VOX-Universidad de Alcalá de Henares, 1995), though deep revision was required in order to perform the conceptual mapping into a single inventory of about 3,000 basic concepts.

Finally, terminals are headed by symbol $ (e.g. $AUCTION_00, $CADAVEROUS_00, $METEORITE_00, $SKYSCRAPER_00, $VARNISH_00 etc). The borderline between basic concepts and terminals is based on their definitory potential to take part in meaning postulates.

As a concluding remark, this three-layer division of the conceptual space responds to the need of defining a core level of knowledge (i.e. basic concepts), that plays a pivotal role between those universal categories which can facilitate ontological interoperatibility (i.e. metaconcepts) and those particular concepts which can grant immediate applicability (i.e. terminals):

> it may be worth attempting to develop complete and coherent core ontologies (…) by means of a combination of top-down and bottom-up approaches, working at levels which are higher than the “particularist” yet lower than the “universalist” (Floridi 1999: 204).

2.2.2. Ontological properties: thematic frames

In FunGramKB, basic and terminal concepts are not stored as atomic symbols but are provided with semantic properties such as the thematic frame and the meaning

postulate. Both of them are conceptual schemata, since they employ concepts –and not words– as the building blocks for the formal description of meaning. Thus, thematic frames as well as meaning postulates become language-independent semantic knowledge representations.

The construction of thematic frames is closely connected to metaconcepts. Strictly speaking, metaconcepts are not concepts but cognitive dimensions, so they are not assigned either thematic frames or meaning postulates. However, metaconcepts are provided with the prototypical participants from which the thematic frames of their subordinate basic/terminal concepts are constructed. Hence, 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. Let us consider the thematic frame of the basic concept +GIVE_00, which belongs to the metaconceptual dimension #TRANSFER:

(1) (x1: +HUMAN_00 ^ +ANIMAL_00)Agent (x2: +CORPUSCULAR_00) Theme (x3)Origin (x4: +HUMAN_00 ^+ANIMAL_00)Goal

This representation includes an Agent (i.e. the entity that transfers another entity to a third entity), a Theme (i.e. the entity that is transferred), an Origin (i.e. the entity from which another entity is transferred) and a Goal (i.e. the entity to which another entity is transferred). Thematic frames can also include those selectional preferences typically involved in the cognitive situation[11]. Thus, thematic frame (1) describes a prototypical cognitive scenario in which "entity$_1$ (Agent), being typically a human or animal, transfers an entity$_2$ (Theme), a corpuscular[12] entity, from one place (Origin) to another (Goal). It should not be forgotten that, although one or more subcategorization frames can be assigned to a single lexical unit, every concept is provided with just one thematic frame.

Participants in the thematic frame of a concept acquire a different interpretation according to the metaconcept to which that concept belongs. Thus, a key requirement for objectivity is to provide thematic roles with accurate definitions according to the location of thematic frames within the metaconceptual level. In this way, the inventory of thematic roles is dramatically minimized while preserving their semantic informativeness. Theme becomes the key role, because its presence is obligatory in any cognitive situation, whereas the remaining participants are defined with reference to that role. In fact, due to the centrality of the Theme role in these conceptual constructs, we opted to call them "thematic" frames[13].

2.2.3. Ontological properties: meaning postulates

Following Velardi *et alii* (1991), the conceptual content of lexical units can be described by means of semantic features or primitives (i.e. conceptual meaning), or through associations with other lexical units in the lexicon (i.e. relational meaning). Strictly speaking, the latter doesn't give a real definition of the lexical unit, but it describes its usage in the language via "meaning relations" with other lexical units.[14]

Most current natural language processing (NLP) systems adopt a relational approach to represent lexical meanings, since it is easier to state associations among lexical units in the way of meaning relations than describing formally the conceptual content of lexical units. However, although large-scale development of deep-semantic resources requires a lot of time, effort and expertise, two main deficiencies in surface semantics can be definitely overcome: its expressive power is dramatically restricted, and redundancy is highly spread through the knowledge base, as has been shown by Periñán-Pascual and Arcas-Túnez (2007b). Consequently, not only is the expressive power of conceptual meanings much more robust, but the management and maintenance of knowledge also becomes more efficient. In addition, even when surface semantics can be sufficient in some NLP systems (e.g. information retrieval or data mining), the construction of a knowledge base which includes meaning definitions guarantees its use for any NLP task, consolidating thus the concept of resource reuse.

These meaning definitions in FunGramKB are expressed in terms of meaning postulates. A meaning postulate is a set of one or more logically connected predications (e_1, e_2... e_n), i.e. conceptual constructs carrying the generic features of concepts.[15] Consider the formal representation of the conceptual content of the basic concept +LEAVE_00 (2), which belongs to the metaconceptual dimension #MOTION:

(2) +(e1: +MOVE_00 (x1)Agent (x2)Theme (x3)Location (x4)Origin (x5)Goal (f1: (e2: +BE_02 (x2)Theme (x4)Location (f2: +IN_00)Position))Condition (f3: (e3: +BE_02 (x5)Theme (x4)Location (f4: +OUT_00)Position)) Condition)

That is, an Agent makes another entity (Theme) move from an Origin to a Goal, providing that the Theme should be located inside the Origin and the Goal should be located outside the Origin.

Moreover, we argue that this type of representation includes rich semantic descriptions that go well beyond those that only capture those aspects of the meaning of a word that are grammatically relevant. Consider the representation of the conceptual basic unit +DECORATE_00:

(3) +(e1: +CHANGE_00 (x1)Theme (x2)Referent (f1: (e2: +BECOME_00 (x2) Theme (x3: +BEAUTIFUL_00)Attribute))Result)

This basic concept, which is a subordinate of +CHANGE_00, has the following definition: an entity transforms another entity with the result that the entity that is transformed becomes more beautiful. In other words, meaning postulates can offer a rich repository of semantic and pragmatic information.

Following this line, an intriguing issue that divides both linguists and language engineers is the granularity of the semantic metalanguage for meaning description, i.e. how fine-grained or coarse-grained the resulting representation should be. Granularity of meaning postulates in FunGramKB is not as fine as that in human-oriented lexicographical

definitions. For instance, the first three senses of *know* in the *Oxford Advanced Learner's Dictionary* (4) have been merged into one single FunGramKB meaning postulate (5):

(4) Know
1. to have information in your mind as a result of experience or because you have learned or been told about it: *The cause of the fire is not yet known.*
2. to realize, understand or be aware of sth: *She knew she was dying.*
3. to feel certain about sth: *I know things will turn out all right.*

(5) +KNOW_00
+((e1: +THINK_00 (x1: +HUMAN_00)Theme (x2)Referent)(e2: +BE_01 (x2)Theme (x3: +TRUE_00 | +POSSIBLE_00)Attribute))

If NLP knowledge bases stored the same number of meanings that paper-based dictionaries have, it would be very difficult to differentiate formally the various senses in polysemous lexical units, not mentioning the dramatic increase of data to be stored and the consequent combinatory explosion that would occur when disambiguating lexically an input text. Thus, FunGramKB meaning postulates are coarse-grained in comparison with standard lexicography. However, they are fine-grained in comparison with the axioms in other formal ontologies.

3. The FunGramKB Lexicon

The FunGramKB lexical model is basically derived from OLIF (Lieske *et alii* 2001; McCormick 2002; McCormick *et alii* 2004) and enhanced with EAGLES/ISLE recommendations (EAGLES 1993, 1996a, 1996b, 1999; Monachini *et alii* 2003; Underwood and Navarretta 1997; Calzolari *et alii* 2001a, 2001b, 2003).

OLIF (Open Lexicon Interchange Format), an XML-compliant standard for lexical/terminological data encoding, was created in the 90s as part of the OTELO (Open Translation Environment for Localization) project, whose primary goal has been the development of interfaces and formats which can help users to share lexical resources within the translation environment (e.g. machine translation, translation memories, terminology databases, and so on). Although OLIF model was chosen as the starting point for the implementation of the FunGramKB lexical level, some parts of this model had to be re-considered in order to make it conform to the FunGramKB architecture (Figure 1).[16] We soon realised that, for example, confining ourselves to OLIF recommendations would not have allowed us to construct full-fledged lexical frames. Therefore, OLIF was modelled with EAGLES/ISLE specifications with the purpose of designing robust computational lexica. EAGLES (The Expert Advisory Group on Language Engineering Standards) was an initiative, sponsored by the European Commission, which aimed to provide recommendations for the standardization of the human-language technology field. More particularly, the Computational Lexicons Interest Group was in charge of analysing the

main practices in lexicographic encoding by comparing computational lexical resources available in European languages. The objective of ISLE (International Standards for Language Engineering), a joint EU-US project initiated in 2000 as an extension of EAGLES work, is to support R&D on human-language technology issues. For instance, the ISLE Computational Lexicon Working Group is committed to the design of MILE (Multilingual ISLE Lexical Entry), a meta-entry model for multilingual lexical information.

Computationally speaking, FunGramKB lexical entries are saved in the form of feature-value data structures formatted in XML. Indeed, XML was chosen as the formal language for knowledge storage since it helps the system to transfer structured data faster, thus facilitating the access to information. Table 1 contains the types of features being present in FunGramKB lexical entries for English and Spanish.[17] The remainder of this section describes each of these features.

	Noun	**Adjective**	**Verb**	**Adverb**
1. Basic				
1.1. Headword	en/sp	en/sp	en/sp	en/sp
1.2. Index	en/sp	en/sp	en/sp	en/sp
1.3. Language	en/sp	en/sp	en/sp	en/sp
2. Morphosyntax				
2.1. Graphical variant	en/sp	en/sp	en/sp	en/sp
2.2. Abbreviation	en/sp	en/sp	en/sp	en/sp
2.3. Phrase constituents: head	en/sp	en/sp	en/sp	en/sp
2.4. Phrase constituents: particle			en	
2.5. Category	en/sp	en/sp	en/sp	en/sp
2.6. Number	en/sp	sp		
2.7. Gender	sp	sp		
2.8. Countability	en/sp			
2.9. Degree		en		en
2.10. Adjectival position		en/sp		
2.11. Verb paradigm and constraints			en/sp	
2.12. Pronominalization			en/sp	
3. LCM Core Grammar				
3.1. Aktionsart			en/sp	
3.2. Lexical template			en/sp	
3.3. Construction			en/sp	
4. Miscellaneous				
6.1. Dialect	en/sp	en/sp	en/sp	en/sp
6.2. Style	en/sp	en/sp	en/sp	en/sp
6.3. Domain	en/sp	en/sp	en/sp	en/sp
6.4. Example	en/sp	en/sp	en/sp	en/sp
6.5. Translation	en/sp	en/sp	en/sp	en/sp

Table 1. *Features in FunGramKB lexical entries.*

3.1. *Basic information*

3.1.1. Headword

With regard to the morphology of headwords, recall that there are three models of computational lexicon: lemma-based, wordform-based and mixed models. In lemma-based lexica, regular inflected forms can be generated by a morphological component provided with inflectional patterns in the form of rules. This type of lexical model, which has been traditionally used in paper-based dictionaries, minimizes redundancy dramatically. On the contrary, in word-form lexica new entries are created for every morphological variant of lexical units. The advantage of the wordform-based model lies in the simplicity to parse input texts. However, this approach presents some drawbacks: redundancy of information, inefficient management of lexica and inability to predict new inflected forms (Lehrberger and Bourbeau 1988; Trost 2003). Finally, the mixed model has lemmas as headwords, but the complete paradigm of inflected forms is embedded in the entry. FunGramKB adopts a lemma-based model, since cases of inflectional morphology are handled by the morphicon, which is made up of two components: morphoRules, which contains a set of regular expression rules, and morphoDB, a database of irregular word-forms. Consequently, inflectional features in OLIF model such as categories <person>, <tense>, <mood> or <aspect> are not pertinent to the FunGramKB lexicon, since they are not used to describe lemmas but to provide grammatical specifications of word forms in the morphicon. On the contrary, lemma-oriented morphological features proposed by OLIF and EAGLES/ISLE are very similar to those found in FunGramKB lexica.

Consequently, the value of <headword>[18] is represented by the canonical orthographic representation of the lexical unit. Unlike standard lexicography, where the various meanings of the headword are grouped inside a single lexical entry, the FunGramKB lexical entry is word-sense-oriented, i.e. an entry is defined as a collection of features linked to a particular sense of the lexical unit. Thus, *country*$_{\text{nation}}$ and *country*$_{\text{countryside}}$ have two different lexical entries, each one linked to a different concept: +COUNTRY_00 and +COUNTRYSIDE_00 respectively.

3.1.2. Index

The value of <index> is a numerical string which serves to arrange the various senses of a lexical unit. At first sight, and since indices are not assigned on a frequency basis, this feature seems to be of little or no importance in one-sense entries. However, the value of <headword> together with that of <index> creates a unique ID for every sense of a lexical unit. In the above example, the short form used to refer to the senses of *country*$_{\text{nation}}$ and *country*$_{\text{countryside}}$ is through tags such as country_01 and country_02 respectively. Moreover, this ID is used to connect senses inter- and intra-linguistically in an unambiguous fashion. For instance, country_01 and state_02 are related each other by means of concept +COUNTRY_00, and country_02 and campiña_01 through concept +COUNTRYSIDE_00.

3.1.3. Language

Feature <language> indicates the language to which the headword belongs.

3.2. *Morphosyntax*

3.2.1. Graphical variants and abbreviations

FunGramKB stores all graphical variants of the headword (e.g. *colour-color, hierba-yerba*) as well as their abbreviations (e.g. *television-TV*) in the same lexical entry.

3.2.2. Phrase constituents

Headwords can be simple, i.e. consisting of a single orthographic word, or otherwise complex. In the latter case, there is a need to state which word within the phrase serves as the head, which is very likely to undergo inflectional morphological phenomena. Moreover, in this case of complex headwords, phrasal verbs must be differentiated from idioms, whose syntactic patterns are more rigid.

3.2.3. Category

FunGramKB lexica store information about open-class lexical units (i.e. nouns, verbs and adjectives). Some adverbs are also included, mainly those involved in the description of the spatio-temporal setting.

3.2.4. Number

Feature <number> allows language engineers to tag nouns and adjectives as dual (e.g. *cat-cats*), *singulare tantum* (e.g. *dust*), *plurale tantum* (e.g. *trousers*), or common (e.g. *species*). Here duality refers to the opposition singular-plural, in which the plural form is constructed by some inflectional rule applied to the singular form. On the contrary, non-dual lexical units express their number by means of syntactic markers. Those plural signifiers which cannot be constructed out of their singular counterparts are stored in the morphicon.

3.2.5. Gender

Feature <gender> presents values such as dual (e.g. *gato*), just masculine (e.g. *mapa*), just feminine (e.g. *plata*), common (e.g. *artista*) and ambiguous (e.g. *mar*). Here duality refers to those cases in which nouns and adjectives take part in the masculine-feminine opposition through morphological markers, whereas non-dual lexical units express their gender by means of syntactic markers. Those feminine forms which cannot be constructed out of their masculine counterparts are stored in the morphicon.

In fact, "gender" is an umbrella concept which covers terms such as "natural gender", "semantic gender" and "formal gender" (Ambadiang, 1999). Thus, a key issue was to determine which type of gender could be automatically inferred and which one should be stated in lexical entries, so the following points were taken into account:

i. Natural gender, which is assigned to animate entities on the basis of their sex, determines the semantic gender of their corresponding lexical units.
ii. In the case of inanimate entities, there is usually no logics of how semantic gender is assigned to nouns.
iii. Formal gender is subject to the use of morphosyntactic markers to define the gender opposition.

The reader can easily conclude that the behaviour of languages towards the various types of gender is dissimilar. In English, gender plays a minor role, since there is no syntactic device of agreement between nouns and adjectives. Here natural gender is the only one worthwhile to mention, since it determines the choice of third-person singular pronouns. However, natural gender can be always derived from the meaning postulate linked to lexical units. Therefore, there is no point to include feature <gender> in English lexical entries. On the contrary, we decided to use this feature as a descriptor of Spanish nouns and adjectives. Although a priority status occurs, i.e. natural gender prevails and formal gender is more relevant than semantic gender, exceptions are so frequent that no rule can be systematically applied. Consequently, gender is ultimately considered a grammatical accident in Spanish (Alarcos Llorach 1994).

3.2.6. Countability

This feature is essential to explain the syntactic behaviour of both English and Spanish nouns, although little attention has been paid by Spanish lexicographers (Bosque, 1999). Countability is linguistically realized by means of morphosyntactic contrasts, particularly affecting subject-verb concordance and the use of some determiners. The analysis of the world model (i.e. the ontology) does not provide enough information as to infer this categorization of nouns, so the value of this feature cannot be automatically derived from the location that the conceptual referents of nouns take in the subontology of entities. In fact, there is nothing in the make-up of things that can explain why some are perceived as mass and others as individual entities. Therefore, the lexicon must set these distinctions, because they form part of our knowledge on language and not of the reality denoted by language (Bosque, 1999). This view is supported by the fact that two languages can categorize conceptually-similar nouns in a different way. For instance, *consejo* and *mueble* are countable in Spanish but their English equivalents *advice* and *furniture* are uncountable.

Although up to six degrees of countability can be found (Downing and Locke 1992), FunGramKB provides just three values for this feature: a noun is always countable, always uncountable, or can sometimes behave as countable and other times as uncountable.

Indeed, dual countability covers cases of "recategorization" (Lyons, 1968), where Spanish is one of the languages which shows more facility for this phenomenon (Bosque, 1999). Psychologically, it is more natural to deal with lexical recategorization by relating countable and non-countable uses of dual nouns in the same lexical entry, instead of creating an entry for a noun denoting a class of objects and another entry for a noun referring to the material or substance from which a well-delimited unit is extracted.

3.2.7. Degree

Since FunGramKB is not ready to find out the number of syllables in lexical units, feature <degree> enables the system to know if comparative and superlative forms of English adjectives are built in an inflectional or periphrastic way. Irregularities are stored in the morphicon. On the contrary, this feature is not pertinent to Spanish adjectives, since most of them are constructed periphrastically. However, those lexical forms taking the organic comparative and superlative (e.g. *mejor, peor, superior, inferior*, etc.) are stored in the morphicon.

3.2.8. Adjetival position

FunGramKB lexica store information about the standard position of adjectives within the phrase, distinguishing three different values for this feature: just attributive, just predicative and attributive-predicative. In the case of Spanish adjectives in attributive function, a further specification is made, regarding the occurrence of adjectives as just premodifiers, just postmodifiers or (pre/post)modifiers.

3.2.9. Verb paradigm and constraints

These features are used to state whether the inflectional paradigm of a verb is regular or irregular as well as any constraint on voice or tense in the paradigm. Although FunGramKB finally relies on the morphicon for the construction of inflectional forms, these features help the system to determine the morphological submodule to be triggered: morphoRules or morphoDB.

3.2.10. Pronominalization

Pronominalization covers those phenomena involving clitic variations of the headword, i.e. reflexivity and reciprocity. On the one hand, the values of <reflexivity> are described as follows:

i. Never reflexive: no reflexive pronoun can be used with the verb, e.g. *parir*.
ii. Always reflexive: a reflexive pronoun is obligatorily cliticised to the verb, e.g. *jactarse*.

iii. Optionally reflexive: the verb can be reflexively marked, but the presence of the reflexive clitic does not involve a shift in the denotative meaning of the verb, e.g. *ir(se)*.
iv. Grammatically reflexive: this contextual variant of *se* is traditionally conceived as a grammatical device which affects the canonical transitivity of predicates according to one of the following conceptual criteria:
    a) transitive verbs can be reflexively marked in order to establish a relationship of identity between two variables of the lexical template: e.g. *Se miró en el espejo*.
    b) transitive verbs can be reflexively marked in order to place the Theme argument into the background, as occurs in cases such as passive, decausative verbs or indeterminate reflexives (Robertson and Turley 2003): e.g. *Muchas pirámides se construyeron en el México antiguo*, *El vaso se rompió*, or *Por aquí se come mucho helado*.

On the other hand, the values of <reciprocity> are described as follows:

v. Never reciprocal: no reciprocal pronoun can be used with the verb, e.g. *beber*.
vi. Grammatically reciprocal: the reflexive pronoun is used to indicate that the people referenced by the plural subject perform the event to each other, e.g. *casar(se)*. The reciprocal construction always involves the detransitivization of the verb.

The word-sense-oriented architecture of FunGramKB lexica conditions the treatment of pronominalization, especially in those cases where the presence of the clitic alters the meaning of the verb. For example, *acordar* and *acordarse* are linked to different concepts (i.e. +AGREE_00 and +REMEMBER_00 respectively), so two different lexical entries are created. This fact determines that *acordar* is never reflexive but *acordarse* is always reflexive.

Reflexivity and reciprocity are underspecified in both OLIF and EAGLES/ISLE lexicographic models. Whereas the OLIF proposal is restricted to values refl and recip for category <synType>, EAGLES/ISLE suggests only two values for <reflexivity> (i.e. refl and no-refl) and two for <clitic> (i.e. clitic and no-clitic). On the contrary, pronominalization in FunGramKB lexica covers a wider phenomenon of reflexive clitic variations of the headword.

3.3. *The LCM Core Grammar*

As preliminarily outlined in Ruiz de Mendoza and Mairal (2007, 2008) and Mairal and Ruiz de Mendoza (2008), the origin of the LCM is to be found in the concern to account for the way meaning construction processes take place at all descriptive levels, in preparedness for syntactic realization. The model thus incorporates meaning dimensions that have a long tradition in pragmatics and discourse analysis, such as

pragmatic implicature, illocution, and discourse coherence. Hence, the LCM recognizes the following four levels (cf. Mairal and Ruiz de Mendoza, 2009):

i. Level 1, or *argumental layer*, accounts for the core grammatical properties of predicates, i.e. the semantic representation of a lexical item.
ii. Level 2, or *implicational layer*, is concerned with meaning captured constructionally and for inferred meaning related to *low-level situational cognitive models* (or *specific scenarios*), which give rise to meaning implications of the kind that has been traditionally handled as part of pragmatics through implicature theory.
iii. Level 3, or *illocutionary layer*, deals with traditional illocutionary force, which we consider a matter of high-level situational models (or generic scenarios).
iv. Level 4, or *discourse layer*, addresses the discourse aspects of the LCM with particular emphasis on cohesion and coherence phenomena. As with level 2, levels 3 and 4 are concerned with both constructional meaning and with meaning obtained through inferential activity

Each level is either subsumed into a higher-level constructional configuration or acts as a linguistic cue for the activation of relevant conceptual structure that yields an implicit meaning derivation. Interpretive activity at all levels is regulated by a number of cognitive constraints. Appendix 1 schematizes the general architecture of the model.

These four different layers are interrelated by two cognitive processes: *subsumption* and *cueing*. However, for the purposes of this paper only the argumental layer concerns us here. This Level 1 deals with the semantic representation of the predicates in a language. In connection with this, lexical entries are represented in terms of lexical templates. A lexical template is an alternative form of lexical representation that is purely decompositional and based on a semantically enhanced notion of logical structures as posited in RRG (cf. Van Valin, 2005). The format of a lexical template consists of two parts: (i) the semantic module, and (ii) the logical representation or *Aktionsart* module, each of which is encoded differently. Here is the basic representational format for a lexical template:

predicate: [SEMANTIC MODULE<qualia>] [AKTIONSART MODULE: RRG'S ASPECTUAL DISTINCTIONS]

The rightmost hand part of the representation includes the inventory of logical structures as developed in RRG. Recall that RRG formulates a verb class adscription system based on the *Aktionsart* distinctions proposed in Vendler (1967), and the decompositional system is a variant of the one proposed in Dowty (1979). Verb classes are divided into *states*, *activities*, *achievements*, *semelfactives*, and *accomplishments*, together with their corresponding causatives (cf. Van Valin, 2005:45). Additionally, a lexical template includes a semantic module that specifies the semantic and pragmatic properties of a predicate, which are in turn formalized by making use of Pustejovsky's

qualia (1991, 1995). Let us consider the following example (cf. Mairal and Ruiz de Mendoza, 2008).

(6) **fathom:**
EVENTSTR: **know'** (x, y)
QUALIASTR: {Q_F: **MANNER** : MagnObstr **think'** (x, y)
Q_T: Culm **know'** (x,y <ALL>)}]

This predicate is a hyponym of *understand* and inherits all the properties from its superordinate, that is, it designates a state structure with a primitive predicate **know**' modified by two arguments (x, y). As an additional distinguishing parameter, this predicate encodes two *qualia*: the *formal* and the *telic* as part of the semantic module. The formal *quale* describes the great difficulty involved in carrying out the process of thinking, i.e. it includes the semantic attributes by means of which *fathom* is semantically distinguished within the larger set of cognition predicates in English. The telic, as encoded in Q_T: Culm **know'** (x,y), specifies the culmination of the process of acquisition of knowledge, that is, the final process of understanding something. The resulting lexical representation does not only encode those aspects of the meaning of a word that are grammatically relevant but also those semantic and pragmatic properties that form part of the meaning representation of a given predicate.

This new theoretical move towards the development of a knowledge base has brought about a reorientation of the epistemological nature of the lexical component. We claim that a lot of information can be inferred from the ontology to the extent that a conceptualist approach is preferred to a lexicalist (cf. Mairal and Periñán, 2009). Recent research has shown that a conceptualist approach offers a very elegant framework to deal with the lexical-conceptual linkage, that is, how the ontology actually interacts with the lexicon. Briefly put, each variable in the lexical template of a lexical unit is uniquely mapped into one and only one participant in the thematic frame of the concept that lexical unit is linked to. After the application of the CLS Constructor algorithm, the system is able to build a CLS for every *Aktionsart* of the lexical unit. Indeed, the FunGramKB model of logical structure is an enhanced version of that presented in RRG: every subcategorised element in the CLS of a lexical unit is referenced through thematic roles to a participant in the thematic frame of the concept to which that lexical unit is linked, and, in turn, every participant in that thematic frame is referenced through co-indexation to a participant in the meaning postulate of that concept.

Semantic data categories in OLIF are not pertinent to FunGramKB lexical entries, since any type of conceptual knowledge must be described in the ontology. However, entries in the lexicon and concepts in the ontology are linked by means of feature <concept> in such a way that (a) lexical entries sharing the same headword are mapped to different concepts and (b) lexical entries sharing the same meaning are mapped to the same concept. According to corollary (b), lexical units linked to the same concept make up a group similar to a synset in WordNet (Miller 1995; Fellbaum 1998). Undoubtedly,

this is the most efficient approach in multilingual computational lexicography, since concept-oriented clusters contribute to minimize redundancy by maximally reducing conceptual proliferation in NLP knowledge bases (Onyshkevych and Nirenburg 1992; Mahesh 1996; EAGLES 1999).

As can be seen in Figure 2, the LCM Core Grammar in the lexicon contains those attributes whose values allow the system to build automatically the CLSs of lexical units.

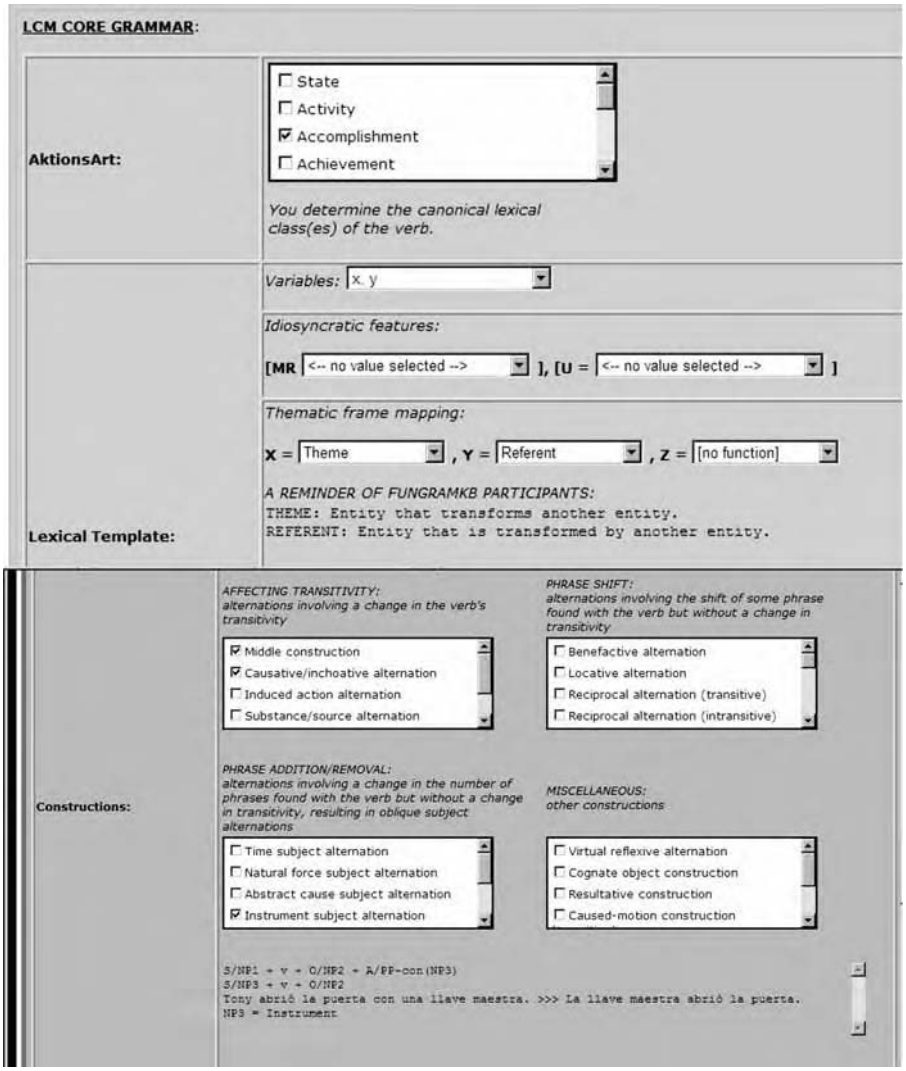

Figure 2. *The LCM Core Grammar in the lexicon interface.*

3.3.1. Syntactic information

The specification of subcategorization frames is the most urgent and complex type of basic linguistic information that an NLP lexicon must provide (EAGLES 1996b). For example, three OLIF data categories are relevant for the construction of these frames:

i. <transType> specifies the type of prototypical transitivity of the verb: e.g. transitive, intransitive, ditransitive, etc.

ii. <synFrame> describes the subcategorization of the lexical entry. A slot-grammar approach is taken for the description of syntactic frames. For example, the frame for the English verb *try* is as follows (McCormick 2002):

    [subj, (dobj-opt | dobj-sent-ing-opt | dobj-sent-inf-opt)]

    In other words, it is a syntactic frame with three optional direct objects realized by a noun phrase, an ing-clause, and an infinitive clause respectively.

iii. <prep> specifies the preposition that fills a "prepositional phrase" slot.

OLIF approach to subcategorization frames presents two main drawbacks. Firstly, OLIF frames are semantically underspecified, since no semantic role is assigned to any slot. Secondly, slot fillers in OLIF are language-specific and not formally represented.

EAGLES/ISLE proposes two types of frame: the syntactic frame, which describes the surface structure, and the semantic frame, which describes the deep structure. The EAGLES/ISLE syntactic (or subcategorization) frame is expressed as a list of slots, where each slot is described in terms of phrasal realization, grammatical function, restricting features and optionality. Moreover, EAGLES/ISLE proposes a FrameSet to be included in the syntactic entry with the aim of collecting surface regular alternations associated with the same deep structure by explicitly linking the slots of the alternating frames by means of rules. Frames involved in a FrameSet are considered to be at the same level, i.e. no alternating frame has a status of privilege from which the other frames are derived through some lexical rule. Surprisingly, the EAGLES/ISLE approach is not as descriptively economical as the traditional approach, where, given two alternating frames, one of them is deemed to be basic and the other derivative.

In our approach, the number of variables is determined from that *Aktionsart* with the highest number of arguments. For example, *freeze* is assigned two variables, those coming from the causative accomplishment class. Following RRG, lexical entries do not include subcategorization features of the arguments (e.g. syntactic function), but just the number of arguments.[19] Since the notion of transitivity is related to the notion of Macrorole[20], the theory distinguishes between "S(yntactic)-Transitivity", which refers to the number of direct core arguments, and "M(acrorole)-Transitivity", which refers to the number of macroroles that a verb allows. Accordingly, the S-transitivity of a verb is less indicative of its grammatical behaviour in simple sentences than its M-transitivity, and consequently, verbs are classified in terms of their M-transitivity. Accordingly, there will be atransitive (macrorole = 0), intransitive (macrorole = 1) and transitive (macrorole =

2) verbs. The presence of idiosyncratic features in the lexical entry implies that the Default Macrorole Assignment Principle is overridden. Some exceptional macrorole assignments are expressed by means of the feature [MR = α], where α can be 0, 1 or 2. This is the case of *belong*, which is exceptional with regard to the Default Principle since it allows the assignment of only one macrorole (i.e. Undergoer), although the verb is associated to two variables. Another kind of lexical idiosyncratic feature can also be specified in the FunGramKB lexical template. For example, it is necessary to specify that the *z* argument of *donate* is the only possible choice for Undergoer, i.e. [U=z], since *donate* does not allow the typical "dative alternation" of three-argument verbs:

(7) Peter donated his gallery to the museum.
*Peter donated the museum his gallery.

In essence, we claim that the codification of the predicate's syntactic confiturations together with the semantic constraints that motivate such syntactic behaviour make our knowledge base rather exceptional, considering that most NLP knowledge bases have been silent about this particular respect.

3.4. *Miscellaneous*

3.4.1. Dialect and Style

Diatopic and diastratic varieties of lexical units are suitably tackled in FunGramKB lexica through features <dialect> and <style> respectively. The relevance of feature <style>, which is ignored by OLIF and EAGLES/ISLE, lies in the fact that lexical registers become one of the reasons for the "evoked meaning" of lexical units, i.e. a type of meaning which is a potential source of variation between cognitive synonyms (Cruse 1986).

3.4.2. Domain

Lexical units can be topically clustered through feature <domain>, whose importance is outstanding in NLP systems such as information retrieval, or tasks such as word sense disambiguation. The FunGramKB inventory of values for this feature actually consists of a subset of forty-eight basic domains from WordNet Domains (Magnini & Cavaglià 2000),[21] a language-independent hierarchy of about two hundred domain lables (e.g. Architecture, Sport, Medicine etc) from which WordNet synsets have been annotated. In turn, WordNet Domains is based on the Dewey Decimal Classification System (Mitchell *et alii* 1996), which is widely used not just by libraries to classify their publications, but also for cataloguing Internet resources.

3.4.3. Examples

Examples from the *British National Corpus* (Davies) and the *Corpus de Referencia del Español Actual* (Real Academia Española) illustrate the meaning of lexical units for English and Spanish respectively.

3.4.4. Translation

Feature <translation> contains lexical units which best serve as default translation equivalents in other languages. The OLIF category <transfer> defines relations between entries from different lexica. In FunGramKB, this category becomes feature <translation>.

4. Conclusion

Recent research in the LCM has developed the design of a lexico-conceptual knowledge base, i.e. FunGramKB. After a brief presentation of some of the most relevant aspects of the ontology, which is the core module of the conceptual level, this paper discusses the anatomy of the lexical component by describing the different sorts of data that form part of a predicate's lexical entry.

Notes

* Correspondence to: Ricardo Mairal Usón. UNED. Dpto. Filologías Extranjeras y sus Lingüísticas. Paseo Senda del Rey, 7. 28040 Madrid (Spain). E-mail: rmairal@flog.uned.es

1. Financial support for this research has been provided by the DGI, Spanish Ministry of Education and Science, grant FFI2008-05035-CO2-01/FILO. The research has been co-financed through FEDER funds. More information about Lexicom may be found at: http://www.lexicom.es.
2. For a more complete description of the conceptual level, we refer to Mairal and Periñán (2009) and Periñan and Mairal (fc).
3. We use the name "FunGramKB Suite" to refer to our knowledge-engineering tool and "FunGramKB" to the resulting knowledge base. Both of them can be browsed in www.fungramkb.com.
4. The current trend in many language engineering projects is not appropriate: *ad hoc* resources are usually developed for a particular NLP application in a particular domain. This *modus operandi* leads to greater efficiency in knowledge representation, but the main drawback is the lack of flexible portability to other domains or tasks because of the inability to meet the new requirements of other applications (Lenci 2000). Since building a large-scale NLP knowledge base is costly in time and effort, it is eagerly recommended to design reusable and updatable resources, so that they can be easily maintained or improved in different projects along the time (Floridi 1999). Thus, multifunctional knowledge bases should integrate any information potentially relevant to any NLP task. The type of knowledge NLP systems require closely depends on the purpose of the applications themselves. For instance, spell checkers require very little lexical information; on the contrary, text understanding systems usually need to process morphological, syntactic, semantic and pragmatic information of lexical units, as well as non-linguistic knowledge from the world model (Nirenburg and Raskin 2004). Therefore, the most reasonable strategy to implement a multifunctional knowledge base is to make it conform to the requirements of the NLP task but letting the system access additional information if necessary.
5. With regard to multilinguality, Aguado de Cea *et alii* (2007) present a thorough revision of current strategies in knowledge-based systems. In this regard, FunGramKB matches the model of knowledge representation in which links between the ontology and language resources are set. As explained in this section (see Figure 1), a lexicon is developed for every language, while the ontology is able to relate lexical units from different languages.
6. English and Spanish are fully supported in the current version of FunGramKB, although we have just begun to work with other languages, i.e. German, French, Italian, Bulgarian and Catalan.
7. In this paper, the term "lexical unit" is used as a synonym of "predicate", i.e. content words to which morphosyntactic and semantic properties are assigned.
8. Terms such as "class", "category" or "semantic type" are often used in ontology engineering to refer to elements such as FunGramKB "concepts". However, we prefer the latter, since it better describes the domain of processing in the two-tier model of our NLP knowledge base, i.e. lexical level and conceptual level.

9. In brief, we shall like to extend FunGramKB ontology module to include terminological subontologies.
10. FunGramKB ontology is actually split into three subontologies, since subsumption (IS-A) is the only taxonomic relation permitted, and therefore each subontology arranges lexical units of a different part of speech: i.e. #ENTITY for nouns, #EVENT for verbs, and #QUALITY for adjectives and some adverbs.
11. Selectional preferences are stated when they can exert some predictive power on the participant. A protocol is currently being developed in order to determine coherent membership criteria for the selectional preferences of participants in thematic frames.
12. The coinage of the concept +CORPUSCULAR_00 was influenced by the SUMO top-level concept CorpuscularObject, which in turn was borrowed from John Sowa's ontology (Niles & Pease 2001b). Thus, following the SUMO definition, our concept refers to "a SelfConnectedObject whose parts have properties that are not shared by the whole".
13. Our approach to thematic roles conflates into one single layer the two different levels posited in RRG: "verb-specific semantic roles" and "thematic relations" (cf. Van Valin, 2005:53), while preserving the notion of macrorole intact.
14. EuroWordNet (Alonge *et alii* 1998, Vossen 1998) is one of the best-known examples of a multilingual "relational" database, which provides elaborate lexical networks by means of semantic relations between *synsets* (or clusters of synonymous words) within every language-dependent wordnet.
15. Periñán-Pascual and Arcas-Túnez (2004) describe the formal grammar of well-formed predications for meaning postulates in FunGramKB. These meaning postulates are partly inspired in Dik's (1989) Functional Grammar and the system of stepwise lexical decomposition.
16. Indeed, one of the advantages of OLIF is the ease of extensibility and customization of its XML-based format to accommodate to the requirements of a project.
17. The "en" and "sp" tags represent English and Spanish languages respectively.
18. In this paper, we adopt the convention of enclosing the names of features between angle brackets, typing the names of lexical units in italics, and tagging concepts with names in block letters.
19. Periñán-Pascual and Mairal (fc.) provide a detailed account of how thematic frames and lexical templates are fully integrated into conceptual logical structures.
20. Note that apart from thematic relations, RRG recognizes another type of semantic function: macroroles. Macroroles are generalizations across different argument types that have significant grammatical consequences. The group of thematic relations that are subjects in transitive active sentences and prepositional complements in passive sentences will be termed "Actors", and those that make up the group of thematic relations that behave as direct objects in active sentences and as subject in passives will be called "Undergoers". One general way to describe these two macroroles is by regarding them as the "logical subject" and the "logical object", respectively. It is also feasible to say that the Actor is the most agent-like argument, and the Undergoer the most patient-like argument. The assignment of macrorole functions to the arguments is conditioned by the argument positions in Logical Structures, according to the Actor-Undergoer Hierarchy (cf. Van Valin, 2005:58). According to this hierarchy, in the Logical Structure of a predicate with two arguments, the leftmost argument will be the Actor and the rightmost one will be the Undergoer. This is the default situation, but there is one marked option of Undergoer assignment in English, and that is when the Undergoer is the first argument of a two-argument state predicate (third position in the scale) and not the second argument (fourth position in the hierarchy). Alternatively, if the verb in a one-place logical structure has an activity predicate, the macrorole is Actor, while if the verb has a non-activity predicate, the macrorole is Undergoer (cf. Van Valin, 2005:63).
21. WordNet Domains can be browsed and downloaded in http://wndomains.itc.it/wordnetdomains.html.

REFERENCES

Aguado de Cea, G., Montiel Ponsoda, E. y Ramos Gargantilla, J. A. 2007. "Multilingualidad en una aplicación basada en el conocimiento". *Procesamiento del Lenguaje Natural* 38. 77-97.

Alarcos Llorach, E. 1994. *Gramática de la Lengua Española*. Madrid: Espasa Calpe.

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

Allen, J. F. y Ferguson, G. 1994. "Actions and events in interval temporal logic". *Journal of Logic and Computation* 4 (5). 531-579.

Ambadiang, T. 1999. ?La flexión nominal. Género y número". Eds., I. Bosque y V. Demonte. *Gramática Descriptiva de la Lengua Española*. Madrid: Espasa Calpe.

Bateman, J. A. 1990. "Upper modeling: a general organization of knowledge for natural language processing". *Workshop on Standards for Knowledge Representation Systems. Santa Barbara*.

Bateman, J. A., Henschel, R. y Rinaldi, F. 1995. "The Generalized Upper Model 2.0". Technical report. Darmstadt: IPSI/GMD.

Beale, S., Nirenburg, S., y Mahesh, K. 1995. "Semantic analysis in the Mikrokosmos machine translation project". *Proceedings of the Symposium on NLP. Bangkok*.

Bosque, I. 1999. "El nombre común". Eds., I. Bosque y V. Demonte. *Gramática Descriptiva de la Lengua Española*. Madrid: Espasa Calpe.

Calzolari, N., Bertagna, F., Lenci, A., y Monachini, M., eds. 2003. *Standards and Best Practice for Multilingual Computational Lexicons and MILE (the Multilingual ISLE Lexical Entry)*. Deliverable D2.2-D3.2. ISLE Computational Lexicon Working Group. [Documento de Internet disponible en http://www.w3.org/2001/sw/ BestPractices/WNET/ISLE_D2.2-D3.2.pdf].

Calzolari, N., Lenci, A., y Zampolli, A. 2001a. "The EAGLES/ISLE computational lexicon working group for multilingual computational lexicons". *Proceedings of The First International Workshop on Multimedia Annotation. Tokyo. Japan*.

Calzolari, N., Lenci, A., y Zampolli, A. 2001b. "International standards for multilingual resource sharing: the ISLE Computational Lexicon Working Group". *Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management*, 15, *Morristown* (*New Jersey*). Association of Computational Linguistics. 71-78.

Calzolari, N., y Picchi, E. 1994. "A lexical workstation: from textual data to structured database". *Computational Approaches to the Lexicon*. Eds., B. T. Sue Atkins y A. Zampolli. Oxford: Oxford University Press. 439-467.

Corcho, O., Fernández López, M. y Gómez Pérez, A. 2001. *Technical Roadmap v. 1.0*, IST-OntoWeb Project, Madrid, Universidad Politécnica de Madrid. [Documento de Internet disponible en http://www.ontoweb.org/download/ deliverables/D11_v1_0.pdf].

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Davies, M. *BYU-BNC: British National Corpus*. Brigham Young University. [Documento de Internet disponible en http://corpus.byu.edu/bnc/].

Downing, A. y Locke, P. 1992. *A University Course in English Grammar*. Londres: Prentice Hall.

Dowty, D. R. 1979. *Word Meaning and Montague Grammar*. Dordrecht: Foris.

EAGLES. 1993. *EAGLES: Computational Lexicons Methodology Task*. EAGLES Document EAG-CLWG-METHOD/B. [Documento de Internet disponible en http://www.ilc.cnr.it/EAGLES96/method/ method.html].

EAGLES. 1996a. *EAGLES: Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages*. EAGLES Document EAG-CLWG-MORPHSYN/R. [Documento de Internet disponible en http://www.ilc.cnr.it/EAGLES96/morphsyn/ morphsyn.html].

EAGLES. 1996b. *EAGLES: Preliminary Recommendations on Subcategorisation*. EAGLES Document EAG-CLWG-SYNLEX/P. [Documento de Internet disponible en http://www.ilc.cnr.it/EAGLES96/ synlex/synlex.html].

EAGLES. 1999. *EAGLES LE3-4244: Preliminary Recommendations on Lexical Semantic Encoding*. Final Report. [Documento de Internet disponible en http://www.ilc.cnr.it/ EAGLES96/ EAGLESLE.PDF].

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge (Massachusetts): MIT Press.

Floridi, L. 1999. *Philosophy and Computing: An Introduction*, London - New York: Routledge.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. y Schneider, L. 2002. "Sweetening Ontologies with DOLCE". Eds., A. Gómez-Pérez, y V. Richard Benjamins. *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web. 13th International Conference. EKAW 2002, 1-4 October. Siguenza. Spain*.

Goldberg, A. 1995. *A Construction Grammar Approach to Argument Structure*, Chicago: University of Chicago Press.

Goldberg, A. 2006. *Constructions at work: the nature of generalization in language.* Oxford: Oxford University Press.

Ide, N. y Véronis, J. 1994a. "Knowledge extraction from machine-readable dictionaries: an evaluation". *Machine Translation and the Lexicon*. Ed., P. Steffens. Springer Verlag. 19-34.

Ide, N. y Véronis, J. 1994b. "Machine readable dictionaries: what have we learned, where do we go?". Eds., N. Calzolari y C. Guo. *Proceedings of the Post-Coling 94 International Workshop on Directions of Lexical Research. Beijing*. 137-146.

Jackendoff, R. S. 1983. *Semantics and Cognition*. Cambridge (Massachusetts): MIT Press.

Lehrberger, J. y Bourbeau, L. 1988. *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation*. Amsterdam-Philadelphia: John Benjamins.

Lenci, A. 2000. "Building an ontology for the lexicon: semantic types and word meaning". *Workshop on Ontology-Based Interpretation of Noun Phrases. Kolding*.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M. y Zampolli, A. 2000. "SIMPLE: A

general framework for the development of multilingual lexicons". *International Journal of Lexicography* 13 (4). 249-263.

Lieske, C., McCormick, S., y Thurmair, G. 2001. "The Open Lexicon Interchange Format (OLIF) comes of age". In *Proceedings of the Machine Translation Summit VIII: Machine Translation in the Information Age, Santiago de Compostela*. 211-216.

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

Magnini, B. y Cavaglià, G. 2000. "Integrating subject field codes into WordNet". *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation. Athens*. 1413-1418.

Mahesh, K. 1996. *Ontology Development for Machine Translation: Ideology and Methodology*, technical report MCCS-96-292, *Computing Research Laboratory, New Mexico State University. Las Cruces*.

Mahesh, K. y Nirenburg, S. 1995. "Semantic classification for practical natural language processing". *The 6th ASIS SIG/CR Classification Research Workshop: An Interdisciplinary Meeting. Chicago*. 79-94.

Mairal Usón, R. y Faber, P. 2007. "Lexical templates within a functional cognitive theory of meaning". *Annual Review of Cognitive Linguistics* 5. 137-172. Amsterdam: John Benjamins.

Mairal Usón, R. y Periñán-Pascual, C. 2009. "Role and Reference Grammar and Ontological Engineering" *Volumen Homenaje a Enrique Alcaraz.* Alicante: Universidad de Alicante.

Mairal Usón, R. y Ruiz de Mendoza Ibáñez, F. J. 2008. "New challenges for lexical representation within the Lexical-Constructional Model (LCM)". *Revista Canaria de Estudios Ingleses* 57. 137-158. Universidad de La Laguna.

Mairal Usón, R. y Ruiz de Mendoza Ibáñez, F. J. 2009. "Levels of description and explanation in meaning construction". *Deconstructing Constructions*. Eds., C. S. Butler y J. Martín Arista. Ámsterdam/ Philadelphia: John Benjamins. 153-198.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N. y Oltramari, A. 2003. "WonderWeb Deliverable D18: Ontology Library". Laboratory for Applied Ontology, ISTC-CNR.

McCormick, S. 2002. *The Structure and Content of the Body of an OLIF v.2.0/2.1*, the OLIF2 Consortium. [Documento de Internet disponible en http://www.olif.net/documents/NewOLIFstruct&content.pdf]

McCormick, S., Lieske, C. y Culum, A. 2004. *OLIF v.2: A Flexible Language Data Standard*, the OLIF2 Consortium. [Documento de Internet disponible en http://www.olif.net/documents/OLIF_Term_Journal.pdf]

Miller, G. A. 1995. "WordNet: a lexical database for English". *Communications of the ACM* 38 (11). 39-41.

Mitchell, J. S., Beall, J., Matthews, W. E. y New, G. R., eds. 1996. *Dewey Decimal Classification*, edition 21. Forest Press, Albany, New York.

Monachini, M., Bertagna, F., Calzolari, N., Underwood, N., y Navarretta, C. 2003. *Towards a Standard for the Creation of Lexica*. ELRA European Language Resources Association. [Documento de Internet disponible en http://www.elra.info/services/standard_lexica.pdf]

Niles, I. y Pease, A. 2001a. "Origins of the Standard Upper Merged Ontology: A *Proposal for the IEEE Standard Upper Ontology"*. *Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology. Seattle*.

Niles, I. Y Pease, A. 2001b. "Towards a standard Upper Ontology". Proceedings of the *Second International Conference on Formal Ontology in Information Systems. Ogunquit*.

Nirenburg, S., Beale, S., Mahesh, K., Onyshkevych, B., Raskin, V., Viegas, E., Wilks, Y. y Zajac, R. 1996. "Lexicons in the MikroKosmos Project". *Proceedings of the AISB'96 Workshop on Multilinguality in the Lexicon. Brighton*.

Nirenburg, S. y Raskin, V. 2004. *Ontological Semantics*. Cambridge (Massachusetts): MIT Press.

Noy, N. F. y McGuinness, D. L. 2001. "Ontology Development 101: A Guide to Creating Your First Ontology". Technical report KSL-01-05, Stanford Knowledge Systems Laboratory, Stanford University.

Onyshkevych, B. A. y Nirenburg, S. 1992. "Lexicon, ontology, and text meaning". *Lexical Semantics and Knowledge Representation. First SIGLEX Workshop, Berkeley.* Eds., J. Pustejovsky y S. Bergler. Berlin-Heidelberg: Springer Verlag. 289-303.

Pedersen, B. S. y Keson, B. 1999. "SIMPLE - Semantic Information for Multifunctional Plurilingual Lexica: some examples of Danish concrete nouns". Proceedings of the SIGLEX-99 Workshop. Maryland.

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

Periñán-Pascual, C. y Arcas-Túnez, F. 2007a. "Deep semantics in an NLP knowledge base". *12th Conference of the Spanish Association for Artificial Intelligence*. 279-288.

Periñán-Pascual, C. y Arcas-Túnez, F. 2007b. "Cognitive modules of an NLP knowledge base for language understanding". *Procesamiento del Lenguaje Natural* 39. 197-204.

Periñán-Pascual, C. y Maizal, R., fc. "Integrating lexico-conceptual knowledge for NLP: the cognitive-lexical linkage".

Procter, P., ed. 1978. "Longman Dictionary of Contemporary English". Harlow (Essex): Longman.

Pustejovsky, J. 1991. "The Generative Lexicon". Computational Linguistics 17 (4). 409-441.

Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge (Massachusetts): MIT Press.

Real Academia Española. *CREA: Corpus de Referencia del Español Actual*. [Documento de Internet disponible en http://corpus.rae.es/creanet.html].

Robertson, J. S. y Turley, J. S. 2003. ?A Peircean analysis of the American-Spanish clitic pronoun system". *Semiotica* 145 (1-4). 21-70.

Ruiz de Mendoza Ibáñez, F. J. y Mairal, R. 2008. "Levels of description and constraining factors in meaning construction: an introduction to the Lexical Constructional Model". *Folia Linguistica* 42 (2). 355-400.

SIMPLE Specification Group. 2000. "Specification SIMPLE Work Package 2". Linguistic Specifications Deliverable D 2.1.

Stockwell, Robert P., Donald Bowen, J. y Martin, J. W. 1965. *The Grammatical Structures of English and Spanish*. Chicago: The University of Chicago Press.

Trost, H. 2003. "Morphology". *The Oxford Handbook of Computational Linguistics*. Ed., R. Mitkov. Oxford: Oxford University Press. 25-47.

Underwood, N. y Navarretta, C. 1997. "Towards a standard for the creation of lexica". Technical report. Center for Sprogteknologi. Copenague.

Van Valin, R. D. Jr. 2005. *The Syntax-Semantics-Pragmatics Interface: An Introduction to Role and Reference Grammar*, Cambridge: Cambridge University Press.

Velardi, P., Pazienza, M. T. y Fasolo, M. 1991. "How to encode semantic knowledge: a method for meaning representation and computer-aided acquisition". Computational Linguistics 17 (2). 153-170.

Vendler, Z. 1967. *Linguistics in Philosophy*. Ithaca, NY: Cornell University Press.

Vossen, P. 2003. "Ontologies". *The Oxford Handbook of Computational Linguistics*. In Ed., R. Mitkov. Oxford: Oxford University Press. 464-482.

VOX-Universidad de Alcalá de Henares. 1995. *Diccionario para la Enseñanza de la Lengua Española*. Barcelona: Bibliograf.

APPENDIX 1

FunGramKB and the Lexical Constructional Model