Revista de Lenguas para Fines Específicos 23.2 (2017), pp. 16-40
http://dx.doi.org/10.20420/rlfe.2017.171
ISSN: 2340-8561

Servicio de Publicaciones y Difusión Científica
Universidad de Las Palmas de Gran Canaria

ARTEMIS: State of the Art and Future Horizons

Ana Díaz Galán[1]
Universidad de La Laguna
Mª del Carmen Fumero Pérez
Universidad de La Laguna

ABSTRACT

The syntactic parser ARTEMIS (Automatically Representing Text Meaning via an Interlingua-based System) is a prototype intended to process natural language within the environment of the Functional Grammar Knowledge Base (FunGramKB) (Periñán-Pascual & Arcas-Túnez, 2010). Different from other parsing devices, ARTEMIS is grounded on two functional linguistic models: Role and Reference Grammar (RRG) (Van Valin & La Polla, 1997; Van Valin, 2005) and the Lexical Constructional Model (LCM) (Ruiz de Mendoza & Mairal Usón, 2008). However, certain adjustments have to be made to both models in order to meet the requirements derived from the fact that ARTEMIS follows the paradigm of Unification Grammars, in such a way that to comply with this paradigm, the GDE (Grammar Development Environment) within ARTEMIS needs to integrate two types of constructs: a catalogue of Attribute-Value Matrixes (AVMs) to describe grammatical units, and a set of production rules (grammatical, lexical and constructional) to allow it to produce a feature based grammar. It is the aim of this paper to give an overview of the investigation carried out so far within ARTEMIS in relation to these two aspects. We will do so by revisiting the literature in relation to the adjustments made to the linguistic models, especially to the RRG, and by reviewing the efforts made to describe the units and design the rules necessary for the parsing of simple sentences in English. Our paper will conclude by pointing at prospective research needed for the completion of this project.

Keywords: Natural Language Processing, ARTEMIS, FunGramKB, computational grammar, Role and Reference Grammar

[1] **Corresponding author** – Facultad de Filología, Campus de Guajara, San Cristobal de La Laguna 38200, Canary Islands, Spain.

Email: mfumero@ull.edu.es

1. Introduction

This paper offers an overview of the state of research contributing to the development of the ARTEMIS (Automatically Representing Text Meaning via an Interlingua-based System) prototype, a syntactic parser intended to process natural language within the environment of the Functional Grammar Knowledge Base (FunGramKB), as described by Periñán-Pascual & Arcas-Túnez (2010). Different from other language processing tools based on stochastic methods, ARTEMIS is a linguistically founded application, both from a semantic and syntactic point of view. Semantically, it draws from the FungramKB nuclear Ontology and, given its syntactic nature, its most relevant characteristic is that it is based on two rigorous functional linguistic theories, namely, Role and Reference Grammar (RRG) (Van Valin & La Polla, 1997; Van Valin, 2005) and the Lexical Constructional Model (LCM) (Ruiz de Mendoza & Mairal Usón, 2008); RRG was not initially designed with a computational orientation, nevertheless, it was adopted as the theoretical linguistic paradigm for ARTEMIS since, as summarized in Periñán-Pascual & Mairal Usón (2009), this lexical-functionalist model presents certain characteristics that make it adequate for the processing of natural languages, namely: it is able to formalize meaning representing it as a logical structure (LS), and its functional approach implies that grammar can only be explained through the interaction of syntax and semantics, a connection which is achieved by means of a linking algorithm. The other functional model which enriches ARTEMIS, the LCM, allows it to account for compositional meaning providing a machine tractable representation of constructions, the constructional schemata.

In two articles published in 2013 and 2014 by Periñán-Pascual and Periñán-Pascual & Arcas-Túnez, respectively, the authors lay out the initial framework of ARTEMIS. They describe the prototype as a bottom-up chart parser with top-down prediction which aims at processing natural language fragments to arrive at their corresponding syntactic and semantic formal representation. The architecture of ARTEMIS has been designed around three main components: the Grammar Development Environment (GDE), which comprises the set of rules (syntactic, constructional and lexical) necessary for the parsing of natural language expressions; the Conceptual Logical Structure (CLS) Constructor, which will produce an initial text meaning representation known as CLS (essentially an evolution of RRG's Logical Structures) and, finally, the COREL-Scheme Builder whose objective is to transform the CLS into the formal FunGramKB representation language (COREL), ultimately arriving at an extended COREL scheme. Such formalization will eventually make ARTEMIS useful for NLP tasks such as information extraction, disambiguation, negation detection, etc. with applications in numerous fields of research.

In the design of ARTEMIS, Periñán-Pascual & Arcas-Túnez (2014) have followed the

paradigm of unification grammars (Boas & Sag, 2012; Sag, Wasow & Bender, 2003). In this line, the GDE -the component where the grammatical, lexical and constructional rules are developed- encapsulates two types of constructs: a catalogue of Attribute-Value Matrixes (AVMs) to describe grammatical units, and the production rules necessary to build a feature based grammar. To date, the major research efforts have focused on the GDE, but neither of these two constructs have yet been fully developed. The aim of this paper is to give an overview of the investigation carried out so far in an attempt to design the rules for parsing simple clauses in English, as well as pointing at prospective research needed for the completion of this NLP project.

2. Adjustments to the linguistic model

ARTEMIS is still in an early stage of development, therefore, the rules it stores, the so called versión 1.0 rules, may be considered preliminary. As Cortés Rodríguez & Mairal-Usón (2016) point out, the computational implementation of RRG as intended by ARTEMIS implies certain adjustments in the linguistic model. Whereas RRG is able to connect the lexical entries of verbs with their syntactic realizations through the linking algorithm, it lacks the mechanisms to satisfactorily account for constructional meaning. Consequently, ARTEMIS needs a constructionist linguistic model, as is the LCM, to be able to process this type of meaning, since, at sentential level, argumental constructions can take precedence over core verbal semantics and, therefore, alter the argumental structure of the predicate, as we will see in the following section. To account for this possibility, Periñán-Pascual & Arcas Túnez (2014) propose a modification of the RRG layered structure of the clause (LSC) by adding a new CONSTR-L1 node between the Clause and the Core nodes. Cortés Rodríguez & Mairal-Usón (2016) further consider that the addition of this new node entails the redefinition of the original RRG Pre-Core slot position as PreC-L1 position. The rationale for this change being that the Pre-Core Slot may house not only those Core constituents stated by RRG (i.e. fronted and interrogative elements), but also constituents which are triggered by a construction, as the following examples proposed by Cortés Rodríguez & Mairal-Usón (2016: p. 96) illustrate:

(1) *For whom* did you wrap the gift? (Beneficiary L1-Construction)
(2) *What* did you open the safe with*?* (Instrumental Construction)
(3) *Into which window* did you kick the ball*?* (Caused-Motion L1-Construction)

Another necessary adjustment to the RRG model derives from the fact that ARTEMIS shares characteristics of unification grammars, in such a way that parsing relies not only on syntactic rules but also on the semantic and grammatical information

contained in the AVMs. Whereas in RRG abstract grammatical categories such as illocutionary force, aspect or negation are described in the operator projection, in ARTEMIS these values -as well as the function words associated with them- are represented in the form of feature bearing matrixes, which now belong to the constituent projection, as illustrated in Figure 1 taken from Cortés Rodríguez & Mairal-Usón (2016, p. 97).

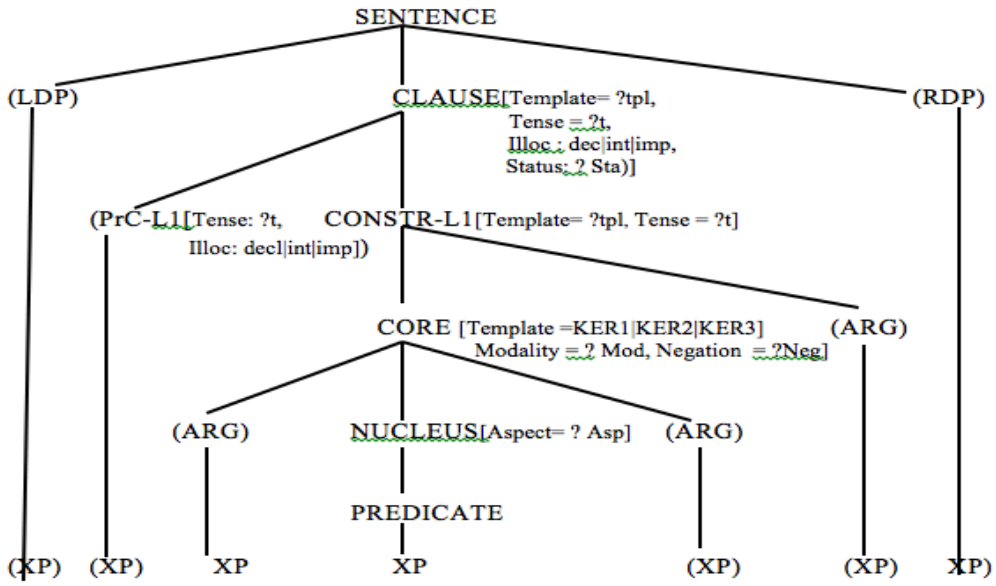


Figure 1. *Modified layered structure of the clause in ARTEMIS (partial).*

As a result of this approach, every grammatical category in the LSC has to be thoroughly described by listing the attributes that define it in a feature bearing matrix. At the same time, these attributes need their own description in another AVM, in such a way that, whenever a category, an attribute or a new part of speech (POS) is introduced, the corresponding AVM must be created and stored in the catalogue of AVMs in the GDE. As a result -with the exception of the Left Detached Position (LDP) and the Right Detached Position (RDP)-, all the other elements of the LSC have been described in the work of various authors. The following list gathers these up to date descriptions:

CLAUSE (Cortés-Rodríguez 2016b, p. 91)

<Category Type="CL">
<Attribute ID="Akt" />
<Attribute ID="Concept " />

<Attribute ID="Illoc" />
<Attribute ID="Status" />
<Attribute ID="Template" />
<Attribute ID="Tense" />
</Category>

PrC-L1 (adapted from Mairal-Usón & Cortés-Rodríguez, 2017)

<Category Type=" PrC-L1">
<Attribute ID="Akt " />
<Attribute ID="Concept " />
<Attribute ID="Illoc" />
<Attribute ID="Mod " />
<Attribute ID="Neg" />
<Attribute ID="Sta" />
<Attribute ID="Template " />
<Attribute ID="Tense" />
<Attribute ID="Weight" />

CONSTR-L1 (adapted from Cortés-Rodríguez 2016b, p. 89)

<Category Type=" CONSTR-L1 ">
<Attribute ID="Akt " />
<Attribute ID="Concept " />
<Attribute ID="Illoc " />
<Attribute ID="Mod" />
<Attribute ID="Sta" />
<Attribute ID="Template" />
<Attribute ID="Tense" />
<Attribute ID="Weight" />
</Category>

CORE (Cortés-Rodríguez 2016b, p. 81)

<Category Type="CORE">
<Attribute ID="Concept " />
<Attribute ID="Illoc " />
<Attribute ID="Mod" />
<Attribute ID="Neg" />
<Attribute ID="Num" />
<Attribute ID="Per" />
<Attribute ID="Recip" />
<Attribute ID="Reflex" />

<Attribute ID="Sta " />
<Attribute ID="Template" />
<Attribute ID="Tense" />
</Category>

NUC (Cortés Rodríguez & Mairal-Usón, 2016, p. 43)

<Category Type="NUC">
<Attribute ID="Asp" />
<Attribute ID="Concept"/>
<Attribute ID="Illoc" />
<Attribute ID="Mod" />
<Attribute ID="Num" />
<Attribute ID="Per" />
<Attribute ID="Recip" />
<Attribute ID="Reflex" />
<Attribute ID="Sta" />
<Attribute ID="Template" />
<Attribute ID="Tense" />
</Category>

PRED (Adapted from Cortés Rodríguez & Mairal-Usón, 2016, p. 22)

<Category Type="PRED">
<Attribute ID="Akt " />
<Attribute ID="Concept"/>
<Attribute ID="Illoc" />
<Attribute ID="Num" />
<Attribute ID="Per" />
<Attribute ID="Recip" />
<Attribute ID="Reflex" />
<Attribute ID="Template" />
<Attribute ID="Tense" />
</Category>

ARG (Adapted from Martín Díaz, this volume)

<Category Type="ARG">
<Attribute ID="Concept" />
<Attribute ID="Macro" />
<Attribute ID="Num" />
<Attribute ID="Per" />
<Attribute ID="Phrase" />

<Attribute ID="Role" />
<Attribute ID="Template" />
<Attribute ID="Variable" />
</Category>

AAJ (adapted from Mairal-Usón & Cortés-Rodríguez, 2017)

<Category Type="AAJ">
<Attribute ID="Concept"/>
<Attribute ID="Macro" />
<Attribute ID="Num" />
<Attribute ID="Per" />
<Attribute ID="Phrase" />
<Attribute ID="Prep" />
<Attribute ID="Role" />
<Attribute ID="Template" />
<Attribute ID="Variable" />
</Category>

ADJUNCT (Adapted from Martín Díaz, this volume)

<Category Type="ADJUNCT">
<Attribute ID="Concept"/>
<Attribute ID="Phrase" />
<Attribute ID="Prep" />
<Attribute ID="Role" />
</Category>

As was pointed out, the attributes that characterize each of the categories above, which would belong in the RRG operator projection, should also be defined by means of AVMs, for example, the AVM for the attribute illocutionary force ("Illoc") has been described as follows:

Illoc (Adapted from Cortés Rodíguez & Mairal-Usón, 2016, p. 100)

<Attribute ID="Illoc" obl="+"num="1">
<Value>?illoc</Value>
<Value Tag= "declarative" >dec</Value>
<Value Tag= "interrogative" >int</Value>
<Value Tag= "imperative" >imp</Value>
</Attribute>

In the same line, an AVM must be designed for every newly created part of speech (POS). For instance, a category such as AUXN, conceived to account for enclitic negative forms of primary auxiliaries, has been described as:

AUXN (adapted from Martín Díaz, this volume)

<Category Type="AUXN">
<Attribute ID="Aspect"/>
<Attribute ID="Illoc"/>
<Attribute ID="Num"/>
<Attribute ID="Per "/>
<Attribute ID="Pol"/>
<Attribute ID="Syn"/>
<Attribute ID="Tense"/>
</Category>

Likewise, the function words corresponding to such categories must also be described and stored in the GDE in the form of lexical rules, as illustrated by the following rule for the past negative and enclitic form of the primary auxiliary DO:

Lexical Rule for *didn't* (Díaz Galán & Fumero Pérez, 2016)
didn't [Pol:n,Tense:past]

The parsing of units smaller than the clause such as the phrase has brought by further adjustments to the RRG linguistic model. Parallel to the changes made to the LSC, Cortés-Rodríguez (2016b) proposes a reinterpretation of the layered structure of Noun Phrases and Adjective Phrases. To maintain consistency with the description of the grammatical components of the clause which is based on functionality, and following Van Valin´s (2008) proposal, the labels Noun Phrase (NP) and Adjective Phrase (AdjP) are replaced with the functionally and typologically motivated labels Referential Phrase (RP) and Modifier Phrase (MP). As stated in Cortés-Rodríguez (2016b, p. 83):

> There is no restriction for RPs to be headed exclusively by any specific lexical category. The same as there is a strong tendency, but not an absolute correlation, for verbs to be the nuclei of clauses, there is a strong tendency for nouns to be the nuclei of RPs, but it is not always the case that there is a nominal nucleus.

In the same article, the author develops the internal layered structure of these two types of phrasal constituents. As we can see in Figure 2 (Cortés-Rodríguez, 2016b, p. 84), the layered structure of phrases also presents constituent projection and operator projection:

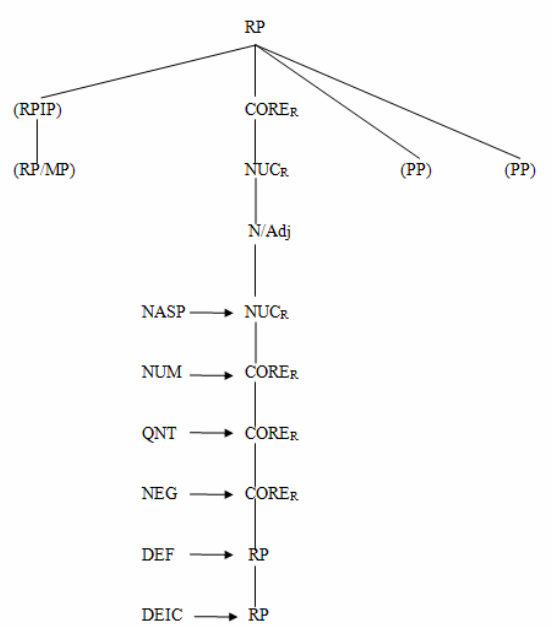


Figure 2. *Layered structure of Referential Phrases.*

RPs include the following layers: (a) a Nucleus, commonly a noun or an adjective; (b) a Core, which integrates both the Nucleus and some optional modifiers (typically PPs), and (c) the higher layer, the RP itself, that houses both the Core and an optional RP Initial Position (RPIP) which parallels both the detached and extra Core slot positions in the Layered Structure of the Clause. RPIP can be filled by genitive modifiers (such as "*this evening's* conference") or *wh*-words of the type "*which* blue skirt"; definiteness operators, possessives, and demonstratives can also occupy this position. Each of these layers can also be modified by operator categories, such as Nominal aspect (Nuclear operator), which refers to count-mass distinction; Number, Quantification and Negation (Core Operators) and Definiteness and Deixis (RP operators). Let us recall that the operator projection will also be replaced by Unification devices affecting the information stored in the AVMs corresponding to the different nodes of the syntactic projection.

There can also be peripheral elements affecting each of the layers in the LSRP, thus, restrictive modifiers (MPs and relative clauses) are treated as nuclear peripheries (e.g. my *good* friend *who lives in Canada*). Setting PPs and MPs are Core peripheral elements (e.g. the devastating earthquake *in San Francisco in 1906*) and non-restrictive modifiers are RP peripheral elements (e.g. the devastating earthquake in San Francisco in 1906, *which killed more than 3000 people*).

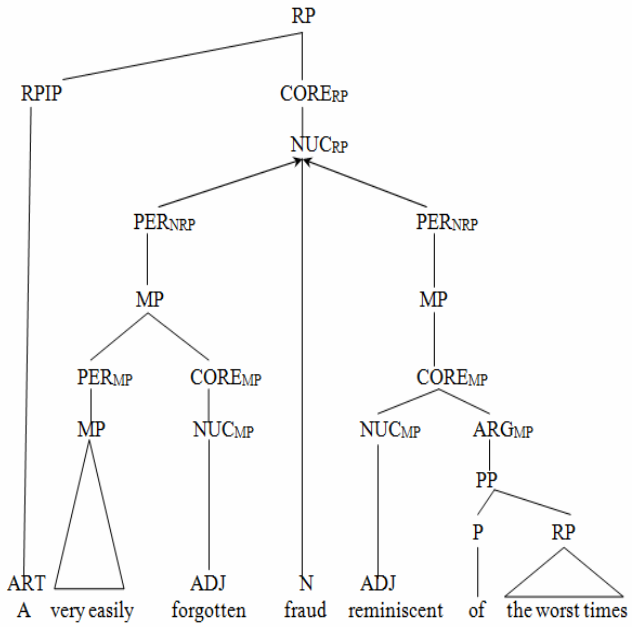
The AVMs for each of the nodes in the LSRP proposed by Cortés-Rodríguez (2016b,

pp. 101-102) are the following[2]:

```
<Category Type="RP">
<Attribute ID="Case" />
<Attribute ID="Count" />
<Attribute ID="Def" />
<Attribute ID="Num" />
<Attribute ID="Prox" />
</Category>

<Category Type="NUC-RP">
<Attribute ID="Count" />
<Attribute ID="Num" />
</Category>

<Category Type="CORE-RP">
<Attribute ID="Count" />
<Attribute ID="Num" />
<Attribute ID="Pol" />
</Category>

<Category Type="RPIP">
<Attribute ID="Def" />
<Attribute ID="Num" />
<Attribute ID="Prox" />
</Category>
```

Finally, modifier phrases also present a parallel Layered Structure (LSMP) as illustrated by the MPs *very easily forgotten* and *reminiscent of the worst times* in the following analysis by Cortés-Rodríguez (2016b, p. 100):

[2] We refer the reader to Cortés-Rodríguez (2016b) for a full-fledged description of phrasal constituents which includes, as well as the AVMs reproduced here, the syntactic rules which govern their behaviour.



3. Accounting for constructional meaning

As we have already pointed out, one of the differences between the linguistic model adopted by ARTEMIS and RRG is that, in the former, constructional meaning is a defining feature. In ARTEMIS, constructional meaning can be derived from the information contained in the core grammar of the verb available in the Lexicon (in the form of lexical templates) and from the constructional schemata contained in the Grammaticon. This module compiles the description of constructions in the form of AVMs which enumerate the main features and establish the set of constraints that characterize each construction. Following the tenets of the LCM, the Grammaticon stores four types of constructions: argumental (L1), implicational (L2), illocutionary (L3) and discursive (L4). However, at its current status of implementation, ARTEMIS can only answer for L1 argumental constructions, that is, those that draw from predicate-argument relations[3].

In an attempt to capture the difference between the information provided by the Lexicon and the Grammaticon, Periñán-Pascual & Arcas-Túnez (2014) make a

[3] For a detailed analysis of how constructional schemata – both for argument and idiomatic constructions – is represented within the GDE, we address the reader to the recent paper by Mairal Usón & Periñán-Pascual (2016).

distinction between Kernel constructions and other argumental constructions. The first are determined by the semantics of the verb, which, depending on the variables in the lexical template stored in the Lexicon, can be: Kernel 0 (zero argument verbs), Kernel 1 (intransitive), Kernel 2 (monotransitive) and Kernel 3 (ditransitive); while the second are those argumental constructions which cannot be derived from the lexical templates. Nevertheless, this distinction between Kernel construction and other argumental constructions may need revision, since, as Luzondo-Oyón & Ruiz de Mendoza (2015) and Fumero Pérez & Díaz Galán (2017) point out, Kernel constructions are not of compositional nature and, therefore, should not be labelled constructions.

Fumero Pérez & Díaz Galán (2017) defend a conception of L1 constructions within FunGramKB that necessarily implies the alteration of the core grammar of the verb for a structure to be considered an L1 construction. Computationally speaking, this distinction is relevant, for it affects the representation of meaning in ARTEMIS, namely the CLS. The changes introduced by a construction may be related to the removal of arguments, the addition of non-optional constituents such as argument adjuncts (AAJs) or secondary nuclei (NUC-s) and/or the modification of the aspectual meaning or *Aktionsart:*

(4) Louise baked a cake *for the kids* (argument adjunct addition).
(5) The pond froze *solid* (secondary nuclei addition).
(6) The window broke (argument removal and change in the inherent *Aktionsart* of the verb *break* from Causative Accomplishment to Accomplishment).

This more restricted view of constructions entails a restructuring of the L1 Constructicon catalogue. Accordingly, in the same paper, Fumero Pérez & Díaz Galán (2017) present a reorganization proposal which distinguishes three main types of argumental constructions: a) constructions which affect the number of arguments, b) those which imply a change in *Aktionsart* and c) constructions which involve both a modification in the number of arguments and in the *Aktionsart*. In turn, Rodríguez–Juárez (in press), after analyzing the behaviour of locative constructions within the framework of RRG, proposes the addition of two new criteria to the previous proposal, namely, d) the L1 construction involves a phrase shift, typically as a result of marked macrorole assignment, and e) the L1 construction changes aspectual meaning (*Aktionsart*) and it also either adds or substracts arguments, or involves a phrase shift, as the following examples illustrate:

(7) Locative Construction: *He spread her toast with butter* (phrase shift as a result of marked macrorole assignment).
(8) Cognate Object Construction: *Sarah sang a song* (change of aspectual

meaning and addition of a constituent).

In the same way that each category in the LSC and every new POS has to be described with an AVM, the semantic and syntactic information that characterizes a construction also needs to be gathered in its corresponding feature bearing matrix to allow the parser to retrieve it and integrate it in the enhanced LSC. This information is stored in the Constructicon within the Grammaticon, which -when completed- will provide a thorough description of the revised L1- Construction catalogue. Figure 3 (proposed by Luzondo-Oyón & Ruíz de Mendoza (2015, p. 17)) exemplifies the type of information contained in the AVM for the Intransitive Motion Construction:

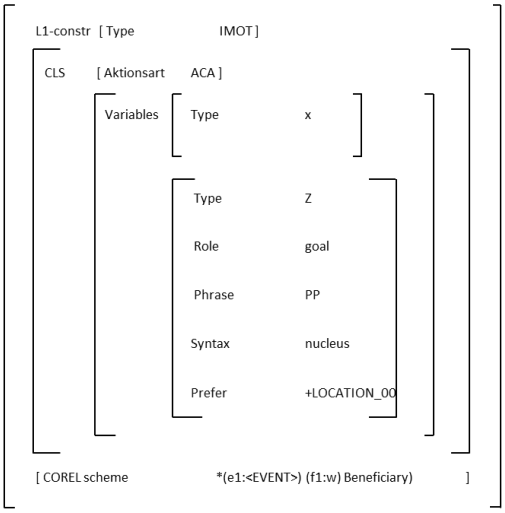


Figure 3. *AVM for the Intransitive Motion construction.*

4. Parsing the simple sentence

Research efforts related to ARTEMIS have so far concentrated on the development of the parsing routine for simple sentences in English. To perform this task we need to provide the parser with three kinds of rules: lexical and constructional rules, as already mentioned, and purely syntactic rules. The former (lexical and constructional rules) "are created in runtime in accordance with the tokens from the input stream" (Periñán-Pascual, 2013, p. 223); the latter, however, have to be designed manually by the linguist.

Syntactic rules allow the parser to distinguish the different realization possibilities of

each of the categories of the LSC. The linguist starts by describing the simplest of the options available for each of the categories, which will become increasingly more complex to account for the different forms the category may adopt. For example, the following rule (Cortés Rodríguez & Mairal-Usón, 2016, p. 107) describes the PRED when it is realized by a single lexical verb:

Rule for the PRED:

PRED[concept= ?, illoc=?, num=?, per=?, tpl=?, t=?]

Likewise, the simplest option for the rule for the NUC would be the following (adapted from Cortés Rodríguez & Mairal Usón, 2016, p. 100):

Rule for the NUC:

NUC [asp=?, concept=?, illoc=?, mod=? num=?, per=?, recip=?, reflex=?, sta=?, tpl=?, =?t=?] → PRED[akt=?, concept= ?, illoc=?, num=?, per=?, recip=?, reflex=?, sta=?, tpl=?, t=?]

These rules for NUC and PRED correspond to simple declarative sentences of the type:

(9) He writes a new chapter.

When the PRED is composed of a lexical verb together with one or more auxiliaries, the rule becomes equally more complex, since it has to render the different combination possibilities, as illustrated in the following rule proposed by Cortés Rodríguez & Mairal-Usón (2016, pp. 100-101):

Rule for the NUC and for the PRED with AUX:

NUC asp=?, concept=?, illoc=?, mod=? num=?, per=?, recip=?, reflex=?, sta=?, tpl=?, =?t=?] → PRED[[akt=?, concept= ?, illoc=?, num=?, per=?, recip=?, reflex=?, sta=?, tpl=?, t=?]] || AUX[asp= pf | pr, illoc= dec | imp , num= pl | sg, per= 1 | 2 | 3, syn= ving | vpar, t= past | pres] PRED[concept= ?, syn= ving | vpar, tpl=?] || MODD[illoc= dec, mod= abl | obl | perm | psbl | vol, num= pl | sg | null, per= 1 | 2 | 3, null, syn= toverb | null, = past | pres | null] PRED[concept= ?, tpl=?] || MODST[illoc= dec, num= pl | sg, | null per= 1 | 2 | 3, | null, sta= inf | nec | poss | subj, syn= toverb | null, t= past | pres] PRED[concept= ?, tpl= ?] || AUX[asp= pf , illoc= dec | imp, num= pl | sg, per= 1 | 2 | 3, syn= apar, t= past | pres] APAR [asp= pr, syn= apar + ving] PRED[concept= ?, syn= ving, tpl=?] || MODD[illoc= dec, mod= abl | obl | perm | psbl | vol, num= pl | sg | null, per= 1 | 2 | 3, null, syn= toverb

| null, t= past | pres | null] AUX [asp= pf | pr, syn= ving | vpar] PRED[concept= ?, syn= ving | vpar, tpl=?] || MODD[illoc=dec, mod= abl | obl | perm | psbl | vol, num: pl | sg | null, per: 1 | 2 | 3 | null, syn= toverb | null, t= past | pres | null] AUX [asp: pf, syn= apar] APAR[asp: pr syn= apar + ving] PRED[concept: ?, syn= ving, tpl=?] || MODST [illoc= dec, num= pl | sg, | null per= 1 | 2 | 3, | null, sta= inf | nec | poss | subj, syn= toverb | null, t= past | pres | null] PRED[concept= ?, syn= toverb | null, tpl=?] ||MODST[illoc= dec, num= pl | sg, | null, per= 1 | 2 | 3, | null, sta: inf | nec | poss | subj, syn= toverb | null, t= past | pres | null] AUX [asp: pf, syn= toverb | null + apar] APAR[asp: pr syn= toverb | null + apar] PRED[concept: ?, syn= ving, tpl:?] ||MODST [illoc= dec, num= pl | sg, | null, per= 1 | 2 | 3, | null, sta= inf | nec | poss | subj, syn= toverb | null, t= past | pres | null] MODD[mod= abl | obl | perm | psbl | vol, syn= toverb | null + toverb] PRED[concept= ?, syn= toverb | null, tpl:?] || MODST [illoc= dec, num= pl | sg, | null, per= 1 | 2 | 3, | null, sta= inf | nec | poss subj, syn= toverb | null, t= past | pres | null] MODD[mod= abl | obl | perm | psbl | vol, syn= toverb | null + toverb] AUX [asp= pf | pr, syn= toverb + vpar | toverb+ ving] PRED[concept= ?, syn= vpar, tpl=?] || MODST [illoc= dec, num= pl | sg, | null, per= 1 | 2 | 3, | null, sta: inf | nec | poss | subj, syn= toverb | null, t: past | pres | null] MODD[mod= abl | obl | perm | psbl | vol, syn= toverb | null + toverb] AUX [asp= pf, syn= toverb+ apar] AUX [asp= pr, syn= apar + ving] PRED[concept= ?, syn= ving, tpl=?]

Cortés Rodríguez & Mairal-Usón (2016) situate all auxiliary verbs, as well as the predicate (PRED), within the NUC node and distinguish the following Nodes to account for the different types of auxiliary verbs: AUX ('Auxiliary verb'), APAR ('Auxiliary verb- past participle'), MODD ('Modal Auxiliary verb - Deontic'), MODST ('Modal Auxiliary verb – Epistemic'). The authors explain that:

> It has been necessary to establish a distinction between the two types of modal verbs as they are the formal realization of different operators in the LSC, namely the CORE operator of Modality and the CLAUSE operator of Status. (Cortés Rodríguez & Mairal-Usón, 2016, p. 102)

The rule for the CORE node (Cortés-Rodríguez, 2016a), representing simple declarative sentences in English, must subsume the three possible syntactic patterns with their different argumental realizations: Kernel-1, Kernel-2 and Kernel-3:

Rule for the CORE in simple declarative sentences:

CORE[concept=?, illoc=?, mod=?, neg=?, num=?, per=?, recip=?, reflex=?, sta=?, tpl=?, t=?] -> ARG[concept=?, macro= A | U | n, num=?, per=?, phrase=?, role=agent | attribute | goal | instrument | location |

manner | origin | referent | result | theme, tpl=?, var= x | y | w | z]
NUC[asp=?, concept=?, illoc=?, mod=?, num=?, per=?, recip=?, reflex=?,
sta=?, tpl=?, t=?] || ARG[concept=?, macro= A | U | n, num=?, per=?,
phrase=?, role: agent | attribute | goal | instrument | location |
manner | origin | referent | result | theme, tpl=?, var= x | y | w | z]
NUC[asp=?, concept=?, illoc=?, mod=?, num=?, per=?, recip=?, reflex=?,
sta=?, tpl=?, t=?] ARG[concept=?, macro= A | U | n, num=?, per=?,
phrase=?, role=agent | attribute | goal | instrument | location |
manner | origin | referent | result | theme, tpl=?, var= x | y | w | z] ||
ARG[concept=?, macro= A | U | n, num=?, per=?, phrase=?, role=agent |
attribute | goal | instrument | location | manner | origin | referent |
result | theme, tpl=?, var= x | y | w | z] NUC[asp=?, concept=?, illoc=?,
mod=?, num=?, per=?, recip=?, reflex=?, sta=?, tpl=?, t=?] ARG[concept=?,
macro= A | U | n, num=?, per=?, phrase=?, role=agent | attribute | goal
| instrument | location | manner | origin | referent | result | theme,
tpl=?, var= x | y | w | z] ARG[concept=?, macro= A | U | n, num=?,
per=?, phrase=?, role=agent | attribute | goal | instrument

The modification of the RRG Layered Structure of the Clause brings by the corresponding alteration of the existing (version 1.0) syntactic rules within ARTEMIS, which now have to incorporate the new nodes: CONSTR-L1, PreC-L1, LDP and RDP. Thus, Cortés-Rodríguez (2016b, pp. 89-90) proposes the following rules for the Sentence, and the Clause which integrate such changes:

S —> CL|| LDP CL|| CL RDP

Rules for Sentence and Clause:

CL [Template=?tpl, Tense = ?t, Illoc : dec|int|imp, Status: ? Sta)] ->
CONSTR-L1[Template= ?tpl, Tense = ?t] || CONSTR-L1[Template= ?tpl,
Tense = ?t] PER || AUX [Tense: ?t, Illoc: int|imp] CONSTR-L1[Template=
?tpl, Tense = ?t] || AUX [Tense: ?t, Illoc: int|imp]-L1[Template= ?tpl,
Tense = ?t] PER || PreC-L1 AUX [Tense: ?t, Illoc: decl|int|imp] CONSTR-
L1[Template= ?tpl, Tense = ?t] || PreC-L1 AUX [Tense: ? t, Illoc: int|imp]
CONSTR-L1[Template= ?tpl, Tense = ?t] PER

4.1. Operations on the simple sentence

To allow ARTEMIS to identify and parse non declarative sentences, it is necessary to provide the GDE with the grammatical information involved in the operations on the simple sentence. In this line, Díaz Galán & Fumero Pérez (2016) describe the use of the DO operator in simple sentences and its role in negation, interrogation and inversion. The first aspect they address in their analysis is the need to expand the

use of the RRG label auxiliary (AUX) so that it may include a number of functional items, that is, the different types of auxiliaries which the parser may find, along with their negative counterparts. This implies the description, in the form of AVMs, of the different attributes that characterize the category AUX: a) it must necessarily indicate tense, number and person; b) to account for the change of illocutionary force involved in interrogative sentences the AVM must contain an attribute for such a feature (Illoc); c) it must also include a polarity attribute (Pol) to explain the presence of AUX in negative sentences; and, d) it will also present an aspect attribute (Aspect), which, for the specific cases in which DO insertion takes place, must always show a zero value. Martín Díaz (this volume) further elaborates on this AVM adding a "Syn" attribute which informs about the possible government requisites of the AUX elements:

```
<Category Type="AUX">
      <Attribute ID="Aspect"/>
      <Attribute ID="Illoc"/>
      <Attribute ID="Num"/>
      <Attribute ID="Per "/>
      <Attribute ID="Pol"/>
      <Attribute ID="Syn"/>
      <Attribute ID="Tense"/>
    </Category>
```

To allow ARTEMIS to identify and parse the AUX element, which in RRG is not part of the constituent projection, the next step is to design a syntactic rule for each of the levels involved in these operations. Since negation, interrogation and inversion affect both the predicate and its arguments, Díaz Galán & Fumero Pérez (2016) propose to locate the AUX in the CORE node.

Building on this proposal, Martín Díaz (this volume) develops the rules which will carry out an effective parsing of yes/no interrogative structures, including, therefore, auxiliaries other than DO. In essence, this involves establishing the AVMs for AUX constituents and the lexical rules, not only of all variants of the English auxiliaries BE, DO and HAVE, but also of the enclitic negative forms which incorporate the negative polarity attribute. Martín Díaz also reevaluates the position of the AUX constituent, which in Cortés-Rodríguez (2016b) was located immediately before the CONSTR-L1 node. Martín Díaz (this volume) proposes to locate it in CORE initial position, since one or other options are indistinguishable in the textual sequence, her preference is justified because it is more in line with the analysis in RRG. Thus the rule for the CORE this author proposes would be as follows:

Rule for the CORE with AUX:

> **CORE** [concept=?, illoc=int, mod=?, neg=?, num=?, per=?, recip=?, reflex=?, sta: ?, tpl=?, t=?] -> AUX[asp: null | pf | pr, illoc: int, num: pl | sg, per:1 | 2 | 3, t: past | pres] ARG[concept=?, macro= A | U | n, num=?, per=?, phrase=?,role:agent | attribute | goal | instrument | location | manner | origin | referent | result | theme, tpl=?, var= x | y | w | z] NUC [asp=?, concept=?, illoc=?, mod=?, num=?, per=?, recip=?, reflex=?, sta=?, tpl=?, t=?] || AUX[asp:null | pf | pr, illoc: int, num: pl | sg, per:1 | 2 | 3, t:past | pres] ARG[concept=?, macro=A | U | n, num=?, per=?, phrase=?, role:agent | attribute | goal | instrument | location | manner | origin | referent | result | theme, tpl=?, var= x | y | w | z] NUC [asp=?, concept=?, illoc=?, mod=?, num=?, per=?, recip=?, reflex=?, sta=?, tpl=?, t=?] ARG[concept=?, macro= A | U | n, num=?, per=?, phrase=?, role: agent | attribute | goal | instrument | location | manner | origin | referent | result | theme, tpl=?, var= x | y | w | z] ?] || AUX[asp:null | pf | pr, illoc: int, num:pl | sg, per: 1 | 2 | 3, t:past | pres] ARG[concept=?, macro=A | U | n, num=?, per=?, phrase=?, role:agent | attribute | goal | instrument | location | manner | origin | referent | result | theme, tpl=?, var= x | y | w | z] NUC [asp=?, concept=?, illoc=?, mod=?, num=?, per=?, recip=?, reflex=?, sta=?, tpl=?, t=?] ARG[concept=?, macro=A | U | n, num=?, per=?, phrase=?, role:agent | attribute | goal | instrument | location | manner | origin | referent | result | theme, tpl=?, var=x | y | w | z] ARG [concept=?, macro=A | U | n, num=?, per=?, phrase=?, role:agent | attribute | goal | instrument | location | manner | origin | referent | result | theme, tpl=?, var= x | y | w | z]

This rule is further expanded to include the options of English yes/no questions with modal verbs (Martín Díaz, this volume). This same rule for the CORE can be adapted to account for negative sentences with enclitic DO by inserting AUXN (negative enclitic auxiliary) in pre-nuclear position. In the case of inversion, DO is triggered by a negative adjunct which appears in initial position. Since the new node CONSTR-L1 has been introduced to the LSC in ARTEMIS, the position of this negative element, which in our first proposal (Díaz Galán & Fumero Pérez, 2016) was located in Pre-CORE position, has been reconsidered and is now located in Pre- CONSTR-L1:

(10) Hardly ever (PreCONSTR-L1) does she bake a cake (CORE) for him (CONSTR-L1).

In the rule for the Core for inversion, the AUX element appears in initial position and before the first argument, thus, it is identical to the rule for the interrogative, the only difference being that the attribute of illocutionary force is now declarative, as is

shown in the rule below:

Rule for the CORE in cases of inversion:

> CORE [concept=?, illoc= dec mod=?, neg=?, num=?, per=?, recip=?, reflex=?, tpl=?, t=?] -> AUX [asp=null, illoc= dec, neg=null, num= pl |sg, per= 1| 2 |3, t= past |pres] ARG[concept=?, macro= A|U |n, num=?, per=?, phrase=?, role: agent |attribute |goal |instrument |location |manner |origin |referent |result |theme, tpl=?, var= x |y |w |z] NUC [asp=null, concept=?, illoc=?, mod=?, num=?, per=?, recip=?, reflex=?, sta=?, tpl=?, t=?] || AUX [asp: null, illoc=dec, neg=null, num= pl |sg, per: 1| 2 |3, t= past |pres] ARG[concept=?, macro= A|U |n, num=?, per=?, phrase=?, role= agent |attribute |goal |instrument |location |manner |origin |referent |result |theme, tpl=?, var= x |y |w |z] NUC [asp=null, concept=?, illoc=?, mod=?, num=?, per=?, recip=?, reflex=?, sta=?, tpl=?, t=?] ARG [concept=?, macro= A |U |n, num=?, per=?, phrase=?, role= agent |attribute |goal |instrument |location |manner |origin |referent |result |theme, tpl=?, var= x |y |w |z] || AUX [asp: null, illoc= dec, neg=null, num: pl |sg, per: 1| 2 |3, t: past |pres] ARG[concept=?, macro= A|U |n, num=?, per=?, phrase=?, role= agent |attribute |goal |instrument |location |manner |origin |referent |result |theme, tpl=?, var= x |y |w |z] NUC [asp=null, concept=?, illoc=?, mod=?, num=?, per=?, recip=?, reflex=?, sta=?, tpl=?, t=?] ARG [concept=?, macro= A |U |n, num=?, per=?, phrase=?, role= agent |attribute |goal |instrument |location |manner |origin |referent |result |theme, tpl=?, var= x |y |w |z] ARG [concept=?, macro= A |U |n, num=?, per=?, phrase=?, role= agent |attribute |goal |instrument |location |manner |origin |referent |result |theme, tpl=?, var= x |y |w |z]

5. Future horizons

So far in this paper we have reviewed the research carried out in regard to the parsing of simple sentences in English within the framework of ARTEMIS. We have summarized the adjustments made to the functional linguistic model (RRG) supporting the parser in order to conform to the needs of the application. Implementing ARTEMIS involves a research effort which includes, among other aspects, the updating of the LSC with the introduction of new nodes; accounting for constructional meaning so that it can be integrated within the GDE; and the design of the initial rules concerning the structure of phrasal constituents, declarative simple clauses, interrogative yes-no questions, and clauses involving operations with do- insertion.

The works where all these proposals have been conveyed share a common

methodological strategy which includes the design both of the AVMs for each of the categories and the attributes described, together with the establishment of the relevant linearization rules. Let us keep in mind that linearization plays a crucial role for the effective parsing of structures, especially, if we consider that the order of constituents is not significant for the definition of the syntactic structures in RRG.

There is still, however, plenty of research to be done for a complete development of the parser. In the short term, the rules for other types of simple sentences (passive, imperative, negative sentences with "not", etc.) have to be designed. The creation of the new rules brings by the need to spell out the details of the AVMs corresponding to the new categories.

Once all the tasks in relation to the simple sentence are fulfilled, it would be necessary to develop the rules for complex sentences, taking into account the theory of nexus and the theory of juncture proposed for these structures within RRG. Briefly, this entails designing rules for the combination of nuclei, cores and clauses, and deciding which type of relation holds between the conjoined layers. Such a relation can be subordination, coordination or co-subordination. In close connection with this last issue is the need to account for the interaction of syntactic rules for complex structures and Level 4 Constructions, as they often involve clausal complexes linked by conjuncts and conjunctions. The effect of constructional interaction with theory of nexus is a pending issue also in the LCM. Needless to say, to complete this endeavour, testing the prototype is absolutely vital to confirm that the parser works as expected.

Article history

Paper received: 22 February 2017
Paper received in revised form and accepted for publication: 28 April 2017

About the authors

Dr. Ana Díaz Galán is a senior lecturer at the Universidad de La Laguna (ULL) where she teaches syntax. She also received her PhD degree from this university in 2001. Her main research interests are discourse analysis, grammar, and lexical relations. Her latest research has focused on computational linguistics and, more specifically, on the processing of grammatical phenomena- as described by the functional theory Role and Reference Grammar- within the framework of the

FunGramKB knowledge base. She has been a member of the following research projects funded by the Spanish Ministry of Science: "Construction of a Core Grammar Spanish-English database within the Lexical Constructional Model" (Project No. FFI2008-05035-C02-02)", "Design of English and Spanish Lexical and Argument-Construction Templates. Applications in Information Retrieval Systems within Multilingual Environments" (Project No. FFI2011-29798-C02-02) and "Development of a virtual laboratory for natural language processing from a functional paradigm" (Project No. FFI2014-53788-C3-1-P). She is also a member of the research groups Neurocog, ReTeLe and Lexicom.

Dr. María del Carmen Fumero Pérez is a senior lecturer in the Philology Department at the University of La Laguna where she also received her Ph.D. degree in English Philology in 2001. Her earlier works, deriving from her doctoral dissertation, focused on pragmatics and academic discourse analysis. More recently, as a member of the Lexicom Research Project, her research has dealt with the interaction between lexis and grammar in functional and cognitive models. Within this line of research, she has participated in the research projects entitled "Construction of a Core Grammar Spanish-English database within the Lexical Constructional Model" (Project No. FFI2008-05035-C02-02)" and "Design of English and Spanish Lexical and Argument-Construction Templates. Applications in Information Retrieval Systems within Multilingual Environments" (Project No.FFI2011-29798-C02-02), both funded by the Spanish Ministry of Science. Her latest works, within the project "Development of a virtual laboratory for natural language processing from a functional paradigm" (Project No. FFI2014-53788-C3-1-P), are related to the field of Natural Language Processing, specifically to the development of NLP tools and their applications.

Acknowledgements

This work has been developed within the framework of the research project "Desarrollo de un laboratorio virtual para el procesamiento computacional de la lengua desde un paradigma funcional" (UNED) FF2014-53788-C3-1-P funded by the Spanish Ministry of Science.

References

Boas, H. & Sag, I. (2012). Sign-Based Construction Grammar. Stanford, Cal.: CSLI Publications.

Cortés-Rodríguez, F. (2016a). Parsing simple clauses within ARTEMIS: The computational treatment of the layered structure of the clause in Role and Reference Grammar. *34th International Conference of AESLA*. Alicante, 14-16, April 2016.

Cortés-Rodríguez, F. (2016b). Towards the computational implementation of RRG. *Círculo, 65*, 75-108.

Cortés Rodríguez, F. & Mairal-Usón, R. (2016). Building an RRG computational grammar. *Onomazein*, *34*, 86-117.

Díaz Galán, A. & Fumero Pérez, M. (2016). Developing parsing rules within ARTEMIS: The case of Do auxiliary insertion. In C. Periñán-Pascual & E. Mestre-Mestre (Eds.), *Understanding meaning and knowledge representation: From theoretical and cognitive linguistics to natural language processing* (pp. 283-302). Cambridge: Cambridge Scholars Publishing.

Ferrari, G. (2004). State of the art in computational linguistics. In P. Van Sterkenburg (Ed.), *Linguistics today: Facing a greater challenge* (pp. 163-186). Amsterdam / Philadelphia: John Benjamins.

Fumero Pérez, M. & Díaz Galán, A. (2017). The Interaction of parsing rules and argument-predicate constructions: Implications for the structure of the Grammaticon in FunGramKB. *Revista de Lingüística y Lenguas Aplicadas 12*, 33-44.

Luzondo-Oyón, A. & Ruiz de Mendoza, F. (2015). Argument structure constructions in a natural language processing environment. *Language Sciences 48*, 70-89.

Mairal-Usón, R. & Cortés-Rodríguez, F. (2017). Automatically representing TExt Meaning via an Interlingua-based System (ARTEMIS). A further step towards the computational representation of RRG. *Journal of Computer-Assisted Linguistic Research 1*(1), 61-68.

Mairal Usón, R. & Periñán-Pascual, C. (2009). The anatomy of the Lexicon within the framework of an NLP knowledge base. *RESLA 22*, 217-244.

Mairal Usón, R. & Periñán-Pascual, C. (2016). Representing constructional schema in the FunGramKB Grammaticon. In J. Fleischhauer, A. Latrouite & R. Osswald (Eds.), *Explorations of the syntax-semantics interface* (pp. 77-108). Düsseldorf: Düsseldorf University Press.

Mairal Usón, R. & Ruiz de Mendoza Ibáñez, F. (2009). Levels of description and explanation in meaning construction. In C. Butler & J. Martín Arista (Eds.), *Deconstructing constructions* (pp. 153–198). Amsterdam / Philadelphia: John Benjamins.

Martín Díaz, M. A. (2017). An account of English yes/no interrogative sentences within ARTEMIS. *Revista de Lenguas para Fines Específicos, 23*(2), 41-62.

Periñán-Pascual, C. (2012). En defensa del procesamiento del lenguaje natural fundamentado en la lingüística teórica. *ONOMÁZEIN, 26*(2), 13-48.

Periñán-Pascual, C. (2013). Towards a model of constructional meaning for natural language understanding. In B. Nolan & E. Diedrichsen (Eds.), *Linking constructions into Functional Linguistics: The role of constructions in RRG grammars* (Studies in Language Series) (pp. 205-230). Amsterdam / Philadelphia: John Benjamins.

Periñán-Pascual, C. & Arcas-Túnez, F. (2010). The Architecture of FungramKB. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (ELRA),* (pp. 2667-2674). Malta: European Language Resources Association.

Periñán-Pascual, C. & Arcas-Túnez, F. (2014). The implementation of the CLS constructor in ARTEMIS. In B. Nolan & C. Periñán-Pascual (Eds.), *Language processing and grammars. The role of functionally oriented computational models* (pp. 164-196) Amsterdam / Philadelphia: John Benjamins.

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

Rodríguez-Juárez, C. On the computational treatment of constructions: the place of locative constructions in a knowledge base. In press.

Ruiz de Mendoza Ibáñez, F. (2013). Meaning construction, meaning interpretation and formal expression in the Lexical Constructional Model. In B. Nolan & E. Diedrichsen (Eds.), *Linking constructions into Functional Linguistics: The role of constructions in RRG Grammars* (Studies in Language Series) (pp. 231-270). Amsterdam / Philadelphia: John Benjamins.

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

Sag, I, Wasow, T. & Bender, E. (2003). Syntactic theory: Formal introduction. Stanford: CSLI Publications.

Van Valin, R. (2005). Exploring the syntax-semantics interface. Cambridge: Cambridge University Press.

Van Valin, R. (2008). RPs and the nature of lexical and syntactic categories in Role and Reference Grammar. In R.D. Van Valin, Jr. (Ed.), *Investigations of the syntax-semantics-pragmatics interface* (pp. 161-178). Amsterdam / Philadelphia: John Benjamins.

Van Valin, R. & LaPolla, R. (1997). Syntax. Cambridge: Cambridge University Press.

Appendix 1: List of abbreviations

AAJ	Argument-adjunct
AdjP	Adjective Phrase
APAR	Auxiliary Verb – past participle
ARG	Argument
ARTEMIS	Automatically Representing Text Meaning via an Inter-lingua-based system
AUX	Auxiliary Verb
AUXN	Enclitic negative primary auxiliary
AVM	Attribute Value Matrix
CL	Clause
CLS	Conceptual Logical Structure
COREL	Conceptual Representation Language
CONSTR-L1	Level 1 Argumental Construction
FunGramKB	Functional Grammar Knowledge Base
GDE	Grammar Development Environment
IMOT	Intransitive Motion Construction
L1	Level 1 (argumental)
LCM	Lexical Constructional Model
LDP	Left Detached Position
LSC	Layered Structure of the Clause
LSMP	Layered Structure of the Modifier Phrase
LSRP	Layered Structure of the Referential Phrase
MODD	Modal Auxiliary verb - Deontic
MODS	Modal Auxiliary verb - Epistemic
MP	Modifier Phrase

NLP	Natural Language Processing
NP	Noun Phrase
NUC	Nucleus
PER	Periphery
PreC-L1	Pre-Construction L1
PRED	Predicate
POS	Part of Speech
PP	Prepositional Phrase
RDP	Right Detached Position
RP	Referential Phrase
RPIP	Referential Phrase Initial Position
RRG	Role and Reference Grammar
S	Sentence