UDC 81'367

PARSING PHRASAL CONSTITUENTS IN ASD-STE100 WITH ARTEMIS*

F.J. Cortés-Rodríguez[1], C. Rodríguez-Juárez[2]
[1]University of La Laguna (Tenerife, Spain)
fcortes@ull.edu.es
[2]University of Las Palmas de Gran Canaria (Las Palmas de Gran Canaria,Spain)
carolina.rodriguez@ulpgc.es

This paper is framed within the field of natural language understanding (NLU) and presents the advances that have been carried out in the parsing device ARTEMIS ("Automatically Representing Text Meaning via an Interlingua-Based System"), which is a NLU prototype designed to obtain the syntactic and semantic representation of linguistic structures and which consists of three submodules: while the CLS Constructor and the COREL Scheme Builder are in charge of providing the semantic structures underlying a language fragment, the Grammar Development Environment (GDE) is responsible for establishing the morphosyntactic makeup of sentences.

In particular, we present the steps that have been taken in the design of the production rules and value added matrixes within the GDE for the analysis of phrasal constituents in relation to the controlled natural language, ASD-STE100, Simplified Technical English.

The adaptation of the GDE components to the requirements of such a simplified English will benefit not only ARTEMIS, offering it a validating platform, but also the users of ASD-STE100, who will obtain a parser adapted to their needs.

Key words: *ARTEMIS, natural language understanding, parsing rules, phrasal constituents, simplified technical English.*

For citation: Cortés-Rodríguez, F. J., & Rodríguez-Juárez, C. (2018). Parsing phrasal constituents in ASD-STE100 with ARTEMIS. *Voprosy Kognitivnoy Lingvistiki*, *3,* 97-109.

DOI: 10.20916/1812-3228-2018-3-97-109

Introduction: setting the linguistic and computational scene of our analysis

This paper forms part of a number of contributions that seek to attain the computational implementation of the Lexical Constructional Model [LCM; Mairal-Usón and Ruiz de Mendoza 2009; Ruiz de Mendoza 2013; Ruiz de Mendoza and Mairal-Usón 2008; Ruiz de Mendoza and Galera 2014], a grammatical model which places itself in the communication-and–cognition tradition as described in Van Valin and LaPolla [1997: 8-15]. The LCM can be described as both a lexically-based and a construction-based grammar which aims at accounting for the relationship between syntax and meaning construction. Thus at the heart of the model lie the so-called lexical and constructional templates. The syntactic and semantic features of predicates are encoded in the format of lexical templates, whereas "a constructional template is a high-level or abstract semantic representation of syntactically relevant meaning elements abstracted away from multiple lower-level representations" [FunGramKB: http://www.lexicom.es/drupal/?q=brief]. Constructional templates are classified in different levels according to their degree of abstractness and schematization: Level 1 constructions are non-idiomatic argument structure characterizations of the types described in Goldberg 1995 and 2006 among many others (e.g. caused-motion constructions, resultatives, middle structures). Level 2 or implicational constructions encode low-level idiomatic meaning of the kind encoded in expressions like *What's X doing Y?* [Kay and Fillmore 1999]; level 3 constructions give rise to conventionalized illocutionary meaning (as in the speech function of requesting in the non-idiomatic expression: *Can you X?*; e.g. *Can you pass me the salt?*). Level 4 constructions are discourse constructions and capture the ways a speaker creates the semantic relations that underlie discourse connectedness; *cause, condition, contrast* and *addition* are some of such relations.

The development of the computational counterpart of the LCM has been the focus of research by several authors [as are Periñán-Pascual 2013; Periñán-Pascual and Arcas-Túnez 2007, 2010 a, 2014; Mairal-Usón and Periñán-Pascual 2009, 2016; Mairal Usón and Ruiz de Mendoza 2009; Díaz Galán and

*Acknowledgements: The research is financially supported by the DGI, Spanish Ministry of Education and Science, project No. FFI2014-53788-C3-1-P.

Fumero Pérez 2017; Fumero Pérez and Díaz Galán 2017; Martín Díaz 2017; Cortés-Rodríguez 2016 a/b; Cortés-Rodríguez and Mairal Usón 2016]. All these works aim at developing several resources (URL: http://www.fungramkb.com/nlp/tools.aspx) devoted to different aspects of language, among which the following are included:

– FunGramKB ("Functional Grammar Knowledge Base"), a knowledge base which holds several modules, where deep and surface semantic information is stored [Periñán-Pascual and Arcas-Túnez, 2007, 2010; Periñán-Pascual and Mairal Usón 2009, 2011; Mairal-Usón and Periñán-Pascual 2009].

– Navigator, a tool to help users retrieve information from both the linguistic (lexicon, and grammaticon) and the ontological modules of FunGramKB.

– ARTEMIS ("Automatically Representing Text Meaning via an Interlingua-Based System"), a natural language processing (NLP) prototype implemented as a parsing device within FunGramKB for the computational treatment of the syntax and semantics of sentences [Periñán-Pascual 2013; Periñán-Pascual and Arcas-Túnez 2014].

The complete implementation of these resources will reveal the computational adequacy of the LCM, and will probably locate it as one of the soundest proposals within the scenario of cognitive-constructional models which are being developed computationally, as are Embodied Construction Grammar [Bergen and Chang 2005], Fluid Construction Grammar [Steels 2004; Steels 2011; Steels and Beule 2006] or Template Construction Grammar [Barrès and Lee 2014]. Even though the proliferation of different computational functional and cognitive grammars can be read as a clear sign of the maturity of such models in the linguistic arena, in computational terms there is still much to be done to consider that they have reached a similar stage of development. In fact, one fundamental prerequisite for a computational grammar to come of age is to be evaluated in different natural language processing (NLP) tasks. However, as has been pointed out in Marques and Beuls [2016: 1137]:

> The evaluation of computational construction grammars is currently not reaching further than proof-of-concept grammar fragments that show how to implement a certain language phenomenon and demonstrate the resulting grammar by means of web demonstrations or its use in a simulation-based robotic environment [Trott et al., 2015].

There are some fundamental reasons to explain the slow pace in the development of these models and their subsequent lack of more extensive evaluation metrics that allow them to be compared within the field of computational grammars [Marques and Beuls 2016: 1137-1138]: firstly, the goal of these models, in consonance with their communicative and cognitive spirit, is not reached once a syntactic and morphological analysis is obtained; these should be an instrumental stage in the achievement of fully-fledged accurate semantic representations of natural language fragments. Secondly, several resources are built manually, at least partially. Thus, the conceptual representations of linguistic and ontological units in FunGramKB are implemented by ontological engineers and computational linguists; and the syntactic structures of clauses and constructions that ARTEMIS will produce are not drawn automatically from existing treebanks; they are also grounded on the rules and feature-matrixes designed by linguists.

Despite these drawbacks, it is necessary to develop at least those proof-of-concept resources that can be used to measure, however partially, the feasibility of such grammars. In this regard, our research can be taken as one step further to implement the ARTEMIS resource, which will be useful in a near future to assess the computational adequacy of the LCM, and of functional and constructional models in general. Specifically, we seek to contribute to the development of ARTEMIS for the analysis of a Controlled Natural Language (CNL), as is ASD-STE100[1], by improving and adapting the existing parsing rules and feature-matrixes in ARTEMIS to the requirements of referential and modifier phrases in this CNL.

Within CNLs, ASDE-STE100 is especially adequate for the development of a prototype version of ARTEMIS, since it is based on English with a number of restrictions on the lexical, syntactic and semantic levels [Kuhn 2014: 136]. Therefore, the scope of the linguistic units that have to be manually devised for morphosyntactic parsing and semantic interpretation is reduced in comparison with those for general English.

The remainder of this paper is organized as follows: Section 2 offers a brief description of the architecture of ARTEMIS, within which special attention is given to the Grammar Development Environment

[1] ASD-STE100 stands for Aero-Space and Defence Industries Association of Europe - Simplified Technical English and is often referred to as Simplified Technical English (STE) or simply Simplified English. This CNL seeks to avoid ambiguity in English language maintenance documentation in the aerospace industry and to provide non-native speakers with texts that are easier to understand [Kuhn 2014: 136].

(GDE) as it is the module where our proposal is to be included. Section 3 provides the linguistic and computational framework of our analysis, since it offers a brief overview of the status of phrasal constituents in general English as described in previous works like Van Valin [2008] and Cortés-Rodríguez [2016a]. Section 4 offers the bulk of our research and concentrates on the adjustments and expansions of the rules offered in the previous section to adapt them to the needs of lexical units and phrases in ASD-STE100. Section 5 provides the rules necessary for the effective parsing of ASD-STE100 Referential and Modifier Phrases within ARTEMIS. Finally, some conclusions are offered in Section 6.

1. A brief description of the architecture of ARTEMIS

ARTEMIS is a proof-of-concept prototype linguistically grounded in two robust grammatical models, LCM and RRG, and obtains the lexical, grammatical and conceptual information that is needed for the generation of the morphosyntactic and semantic representation of language fragments from the Lexicon, the Grammaticon for Constructional structures and the Ontology in FunGramKB [Cortés-Rodríguez and Mairal-Usón 2016: 90; Mairal-Usón and Periñán-Pascual 2016: 87].

Within ARTEMIS, the process involved in understanding a stretch of natural language and binding it with their corresponding grammatical and semantic structures can be summarized as follows. In the first place, it is necessary to make an effective computational parsing of the morphosyntactic structure underlying sentences based on the principles of RRG and the LCM for grammatical descriptions [Cortés, 2016a: 80]. This task is done in the Grammar Development Environment (GDE), which comprises the rules that are necessary for the computational parsing of the morphosyntactic structure of sentences. Then, and in order to transfer the shallow semantic representations of sentences into conceptually deeper structures, two other components have been designed: the Conceptual Logical Structure (CLS) Constructor, which will produce an initial text meaning representation that is an enriched version of RRG's Logical Structures, and the COREL-Scheme Builder, which transforms the CLS into COREL (COnceptual REpresentation LAnguage), the formal FunGramKB representation language that formalizes conceptual knowledge in FunGramKB [Cortés 2016a: 80; Díaz Galán and Fumero Pérez 2017]. Figure 1 shows the architecture of ARTEMIS and the process that is followed in understanding a fragment of natural language together with the tools that are activated in each phase of the process (Figure 1).

This research is restricted to the GDE component, where we can distinguish two types of theoretical constructs. The first one comprises the production rules, i.e. the set of lexical, syntactic and constructional rules that are necessary to account for syntactic structures and that will parse the language fragment and generate a morphosyntactic tree. Only the syntactic rules have to be predefined in the GDE, since the lexical and the constructional rules are constructed automatically. The second component is made up of a catalogue or library of Attribute-Value Matrixes (AVMs) that are feature-bearing structures that encode the grammatical features that modify the different categories or units, and that cannot be retrieved from the information that is stored in the Lexicon, the Grammaticon and the Ontology of the knowledge base [Cortes, 2016a: 80-81; Cortés-Rodríguez and Mairal-Usón 2016: 90, 97]. Figure 2 illustrates the behaviour of the GDE and the CLS constructor as presented in Periñán-Pascual and Arcas-Túnez [2014: 178] and Cortés-Rodríguez and Mairal-Usón [2016: 91] (Figure 2).

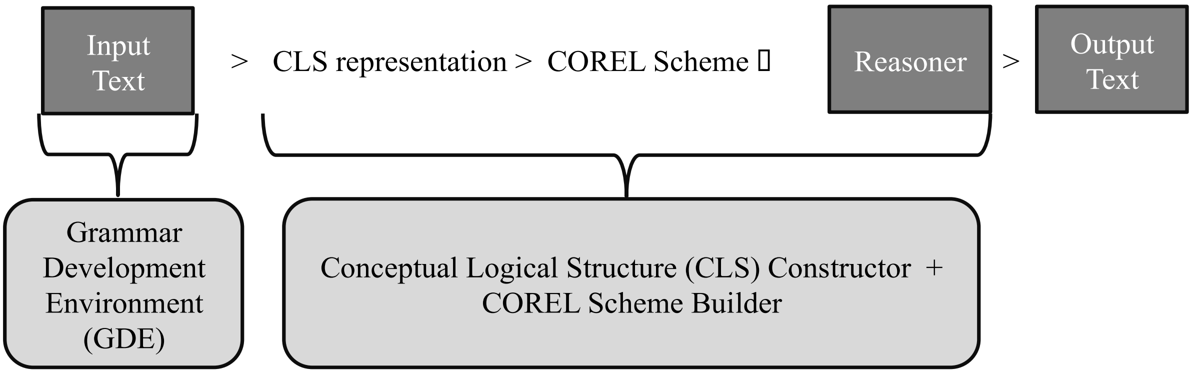

Input Text
> CLS representation > COREL Scheme
Reasoner
>
Output Text
Grammar Development Environment (GDE)
Conceptual Logical Structure (CLS) Constructor + COREL Scheme Builder

Figure 1. The architecture of ARTEMIS

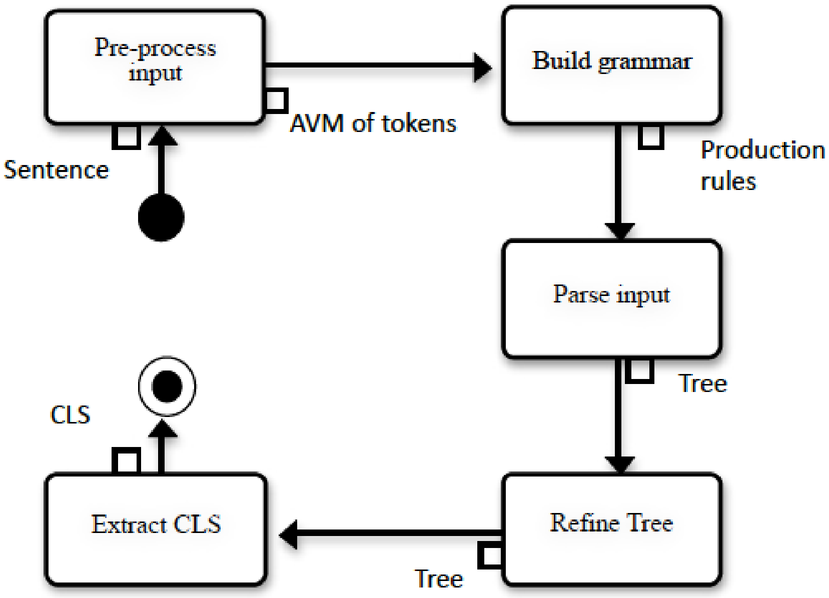


Figure 2. An abridged version of the ARTEMIS process

2. Phrasal constituents in ARTEMIS: a brief overview

Since this research is restricted to the design of the syntactic rules for the grammatical analysis of phrasal constituents in Simplified Technical English, this section presents an overview of the status of this type of constituents taking into account Van Valin's latest proposal [2008] as well as the work by Cortés-Rodríguez [2016a].

Based on Van Valin's assumption that the most significant lexical and syntactic categories are projections of the functional status of the clause constituents and not projections of a lexical head, labels such as noun phrases (NP) and adjectival phrases (AdjP), which are not functionally motivated, have been replaced by two types of constituents, referential phrases (RPs) and modifier phrases (MPs), which are functionally and typologically oriented.

Thus, Cortés-Rodríguez [2016a] presents the Layered Structure of Referential Phrases (LSRP) (see Figure 3, in which, as in the case of clauses, each layer can be modified by different types of operators which introduce grammatical information, such as nuclear operators (nominal aspect (NASP): count-mass distinction), core operators (number (NUM), quantification (QNT), and negation (NEG)) and RP operators (definiteness (DEF), deixis (DEIC)).

There is also a periphery for each layer, as happens in clauses as well, which can be filled in by lexical (*big* in *The three big bridges*) or phrasal constituents (such as prepositional phrases (PPs): *The construction of the bridge by the company in New York City*) that are going to modify the content of their respective layers. Thus, at the level of the Nucleus (NUC) the inner-most nuclear periphery may be subsumed by restrictive MPs (*big*) or restrictive relative clause (*My dear old wood hammer that never lets me down*); at the level of the Core, the core periphery may include setting MPs (*tomorrow*) and PPs (*in New York city*), and at the level of the RP, the periphery node can be occupied by non-restrictive constituents such as non-restrictive relative clauses and appositions (*Rebeca, a cupcake expert*)[1].

In Figure 4, we present the improved rules for phrasal constituents as presented in Cortés-Rodríguez [2016a: 93-94] that show the type of syntactic information that must be incorporated in the repository of syntactic rules within the GDE for the effective parsing of these phrasal constituents at Level 1 and which predict the layered internal configuration of RPs and MPs. For example, the second subset of rules spells out the internal configuration of RPs where we can read that the nucleus of an RP can be realized by a noun (N), or an adjective (ADJ) or a pronoun of different kinds (PROD: pronoun (demonstrative); PROQ: pronoun (quantifier) or a numeral (NUMC: numeral (cardinal); NUMO: numeral (ordinal), among other possibilities.

[1] Examples are taken from Van Valin [2005: 25] and Cortés-Rodríguez [2016a: 86].

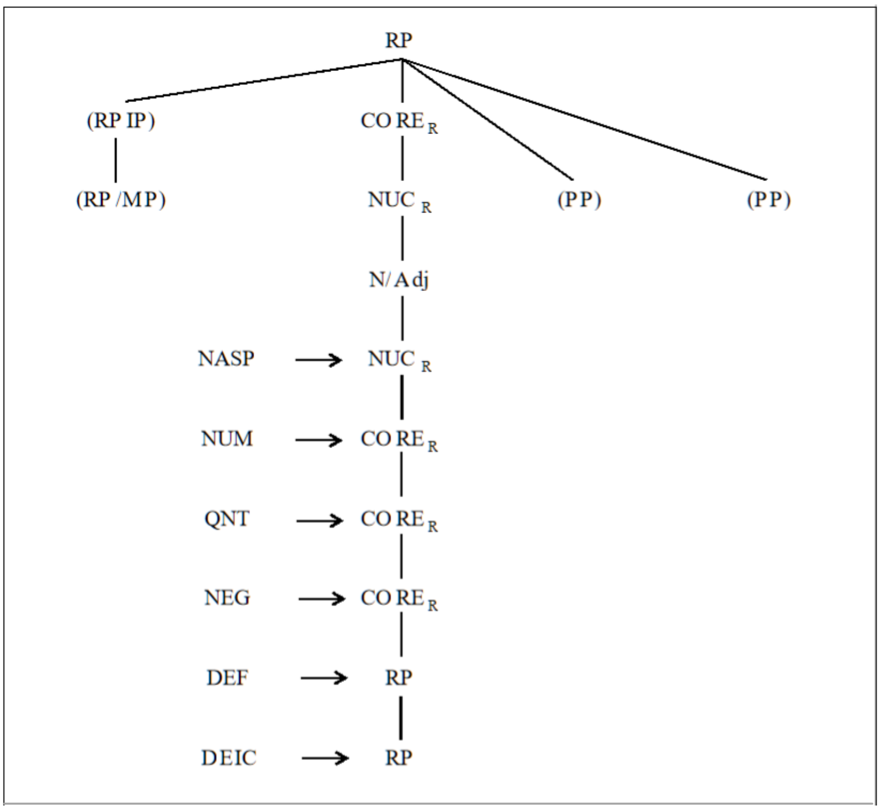
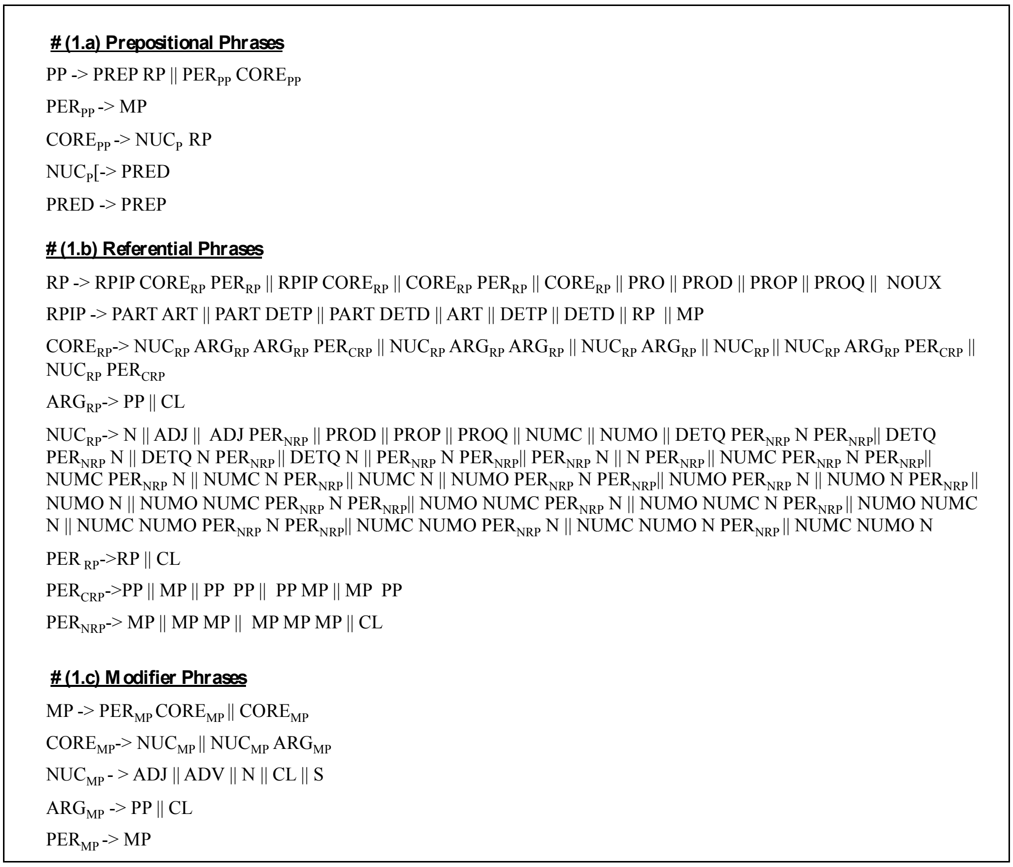


Figure 3. The Layered Structure of Referential Phrases (LSRP) [Cortés-Rodríguez, 2016a: 84]



# (1.a) Prepositional Phrases

PP -> PREP RP || PER$_{PP}$ CORE$_{PP}$

PER$_{PP}$ -> MP

CORE$_{PP}$ -> NUC$_P$ RP

NUC$_P$[-> PRED

PRED -> PREP

# (1.b) Referential Phrases

RP -> RPIP CORE$_{RP}$ PER$_{RP}$ || RPIP CORE$_{RP}$ || CORE$_{RP}$ PER$_{RP}$ || CORE$_{RP}$ || PRO || PROD || PROP || PROQ || NOUX

RPIP -> PART ART || PART DETP || PART DETD || ART || DETP || DETD || RP || MP

CORE$_{RP}$-> NUC$_{RP}$ ARG$_{RP}$ ARG$_{RP}$ PER$_{CRP}$ || NUC$_{RP}$ ARG$_{RP}$ ARG$_{RP}$ || NUC$_{RP}$ ARG$_{RP}$ || NUC$_{RP}$ || NUC$_{RP}$ ARG$_{RP}$ PER$_{CRP}$ || NUC$_{RP}$ PER$_{CRP}$

ARG$_{RP}$-> PP || CL

NUC$_{RP}$-> N || ADJ || ADJ PER$_{NRP}$ || PROD || PROP || PROQ || NUMC || NUMO || DETQ PER$_{NRP}$ N PER$_{NRP}$|| DETQ PER$_{NRP}$ N || DETQ N PER$_{NRP}$ || DETQ N || PER$_{NRP}$ N PER$_{NRP}$|| PER$_{NRP}$ N || N PER$_{NRP}$ || NUMC PER$_{NRP}$ N PER$_{NRP}$|| NUMC PER$_{NRP}$ N || NUMC N PER$_{NRP}$ || NUMC N || NUMO PER$_{NRP}$ N PER$_{NRP}$|| NUMO PER$_{NRP}$ N || NUMO N PER$_{NRP}$ || NUMO N || NUMO NUMC PER$_{NRP}$ N PER$_{NRP}$|| NUMO NUMC PER$_{NRP}$ N || NUMO NUMC N PER$_{NRP}$ || NUMO NUMC N || NUMC NUMO PER$_{NRP}$ N PER$_{NRP}$|| NUMC NUMO PER$_{NRP}$ N || NUMC NUMO N PER$_{NRP}$ || NUMC NUMO N

PER $_{RP}$->RP || CL

PER$_{CRP}$->PP || MP || PP PP || PP MP || MP PP

PER$_{NRP}$-> MP || MP MP || MP MP MP || CL

# (1.c) Modifier Phrases

MP -> PER$_{MP}$ CORE$_{MP}$ || CORE$_{MP}$

CORE$_{MP}$-> NUC$_{MP}$ || NUC$_{MP}$ ARG$_{MP}$

NUC$_{MP}$ - > ADJ || ADV || N || CL || S

ARG$_{MP}$ -> PP || CL

PER$_{MP}$ -> MP

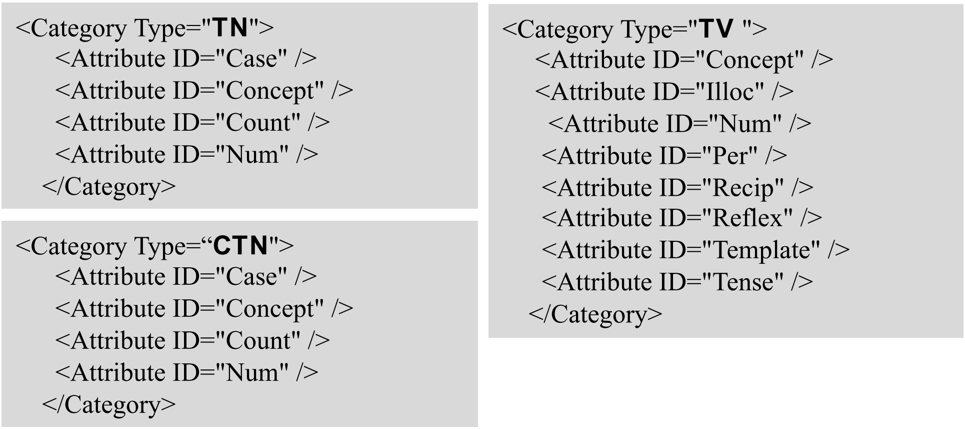
Figure 4. Level 1: Phrasal structures [Cortés-Rodríguez 2016a: 93-94]

3. The treatment of phrasal constituents in ASD-STE100

Our attempt to implement the ARTEMIS resource for the analysis of phrasal constituents in a controlled natural language, as is Simplified Technical English (ASD-STE100) has necessarily implied the adaptation of the syntactic rules presented in Section 3 to the requirements and needs of the lexical units and phrases in this Simplified English. In this process, the most radical variation has been the need to account for the complex and fully productive word-formation processes that are described in the specification document for ASD-STE100 [January 2017].

Despite its name, ASD-STE100 is a simplified language only at sentence level (“you must write short sentences and use simple sentence structure” [Specification 2017: p. 1-4-1]), since a revision of Sections 1 to 3 of Part 1 in the Specification document has revealed the complexity of lexical units and their phrasal projections in ASD-STE100. This complexity derives from the fact that, since RPs are by nature the natural means to provide labels and descriptions of any kind of object or component, this technical language deals with a huge amount of nomenclature (e.g. *main fuel metering unit, distribution block* [Specification, 2017: p. 1-2-3]) and allows the use of a catenation of nouns (*retraction-winch handle*) or adjectives and nouns (*main-gear-door*), and even some other possible combinations (*on-ground configuration, safetied-for-maintenance configuration*) to designate the components of aeroplanes. Additionally, STE permits the use of company-specific or project oriented technical words, in particular technical names (TN) and technical verbs (TV) [Specification 2007: p. ii] that are not included in the controlled dictionary. Thus, within ASD-STE100, we can distinguish two types of lexical units: (i) a restricted vocabulary, constituted by nouns, adjectives, adverbs and verbs, together with the function words present in the Dictionary (Part 2 of the Specification document); and (ii) an unrestricted vocabulary made up of Technical Nouns (TNs, e.g. *cap, engine*), Technical Verbs (TVs, e.g. *ream, dry-motor, wet-motor*), and also Deverbal ‘adjectives’ (-ed and –ing participles functioning as modifiers such as *reamed hole*), of which there is not a closed list (which means that there is no source to check whether a (sequence of) word(s) is a TN) since each manufacturer uses their own technical lexical units. Furthermore, nominal composition allows for any free combinations of both Nouns and Technical Names (and TNs can be created out of other TNs or Ns), and yields as output what we have called Compound Technical Names (CTNs, e.g. *weight-on-wheels condition*). All this contributes to the complexity registered at phrasal level which obviously poses a problem when it comes to determining how to process these constituents in ARTEMIS.

The first adaptation that has been done within the GDE is the registration of three new types of parts of speech (POS) in the catalogue of AVMs in the format of attributes and values (Figure 5), which will have to be later integrated in the relevant positions in the appropriate syntactic rules. In the case of TN and TV, their AVMs will include the same type of Attributes as those of N and V; the AVM for CTN will have the same attributes as the rest of nominal POS (i.e. Ns and TNs):



```
<Category Type="TN">
    <Attribute ID="Case" />
    <Attribute ID="Concept" />
    <Attribute ID="Count" />
    <Attribute ID="Num" />
  </Category>

<Category Type="“CTN">
    <Attribute ID="Case" />
    <Attribute ID="Concept" />
    <Attribute ID="Count" />
    <Attribute ID="Num" />
  </Category>

<Category Type="TV ">
   <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>
```

Figure 5. AVMs for Technical Noun (TN), Technical Verb (TV) and Compound Technical Noun (CTN) in the GDE (ARTEMIS)

Another issue that we have had to address is related to the number of possible lexical units that can be stacked together, since despite there apparently being some restrictions in this number, a thorough revision of the documents has revealed that this is far from being coherently applied. Thus, for instance, one difficulty for parsing is the fact that, when introduced for the first time in a text, TNs can consist of several words, i.e., in their first occurrence, there are limitless possibilities in terms of the number of n-grams, as for instance *ramp service door safety connector pin* [Specification 2007: p.1-2-2], which includes 6-grams. Since there is no lexicon for technical words, it seems necessary to pre-process the documents and build a satellite ontology prior to activating parsing operations. However, even so, there are still some decisions to be taken as to what pertains to the lexical level and what is syntactic in these cases. That is, should the ontology consider *ramp service door safety connector pin* a single lexical unit or take only the last n-gram (plus some immediate preceding n-grams) as a TN? If the last option is taken, the rest should be treated as syntactic and integrated in the GDE. We propose a compromise solution and consider that underived TNs consist prototypically of up to 3 unigrams in a 3-gram sequence, and will allow for them to have more n-grams only if, when introduced for the first time in a document, they are followed by their corresponding acronym between brackets. Thus, in the following extract from Airbus corpus[1], the underlined sequences should be encoded as basic TNs in the Satellite Ontology since they are followed by their corresponding acronyms:

General

The Landing Gear (L/G) (bigram plus acronym) *has a twin-wheel nose gear and a three-strut twin-wheel main gear on the left and right sides. All gears include an oleo-pneumatic shock absorber.*

The L/G includes:

The Nose Landing Gear (NLG) (trigram plus acronym)

The Main Landing Gear (MLG) (trigram plus acronym)

The Landing Gear Extension and Retraction System (LGERS) (5-gram plus acronym)

The steering system

The kneeling system.

(DMC-AJ-A-32-00-00-0AA0-030A-A_15-00)

However, the sequence *Nose Landing Gear doors (NLG doors)* from the same document will be analysed through our syntactic rules in the GDE, as represented below (1) in bracketing: the head of the RP, i.e. its NUC, is the TN *doors* (note that it is not a noun since it is not included in the Dictionary), therefore, it is the NUC-RP which in turn is modified by the Peripheral MP *Nose Landing Gear*, which is a 3-gram TN; both units together form the CORE of the RP; the CORE is complemented by a noun cluster between brackets which provides another means to refer to the combination of both TNs, and should be analysed as an appositive Peripheral RP.

1. *[[[[Nose Landing Gear$_{TN}$]$_{MP}$]PER-NRP[doors$_{TN}$]NUC-RP]CORE-RP] [(NLGdoors$_{TN}$)Apposition] PER-RP]RP*

The analysis of the 3-gram TN *Nose Landing Gear* as an MP is consistent with the restrictions on the functional status of TNs in STE documents: "Use a technical name only as a noun or as an adjective that is part of a technical name. Do not use the same word as a verb" [Specification 2007: p. 1-1-8]. Here 'adjective' must be better understood as 'having a Modifier function' (i.e. we will treat these as cases in which the head of a MP is a TN).

Once a long underived TN is used for the first time, the Specification document includes some rules for noun clustering that must be followed for all subsequent occurrences of such a TN. These rules, which also apply for the creation of derived TNs[2], are explained below:

(a) *Write noun clusters of no more than 3 words* [Rule 2.1, Specification 2007: p. 1-2-1], so, if longer in their first occurrence, users of STE are instructed to reduce them to a 3-gram sequence maximum; the rest of constituents can be reanalysed as MPs (postmodifier PPs introduced by prepositions such as *of, on, in* and *for* (which do not count as words in a noun cluster) and postmodifier clauses) in the RP. For instance, a noun cluster like *the forward turbine overheat thermocouple terminal tags* can be rewritten and reduced to *the terminal tags on the forward overheat thermocouple of the engine* [Specifica-

[1] For courtesy of Airbus in Seville, we have been able to use their corpus of a selection of texts from aircraft maintenance to provide real examples of ASD-STE, which are clearly identified in the text with the Airbus corpus reference in brackets.

[2] Things are not so clear-cut when documents in STE are revised; we have found cases of long noun clusters without any indication of how they are to be reduced on subsequent occasions; e.g. "Remove safety locks (*Landing gear safety pin removal*)". In our proposal for TN identification in the process of ontology building, there is *a priori* no guarantee that a 3-gram sequence like *Landing gear safety* is not automatically analyzed as a TN, unless a restriction is given to prefer as TNs the last 3-gram sequence of longer sequences.

tion 2007: p. 1-2-2], where the 2-word noun cluster *terminal tags* is postmodified by two prepositional phrases.

(b) *Use hyphens (-) between words that are used* ***as a single unit*** [Rule 2.2, Specification 2007: p. 1-2-2] (our emphasis). This rule permits the reduction of long TNs by using hyphenation, although there are some conditions that have to be met: (i) you can hyphenate only up to 3-gram sequences; and (ii) hyphenated words count as one word. As can be seen, hyphenation is *de facto* a very productive compounding strategy in ASD-STE100 (note our emphasis in the quotation above), and shows no restrictions as to the kind of words that can be linked, as illustrated in the following examples:

2. *A two-position PARK BRK switch with a* *pull-to-turn* *mechanism for manual control of the parking brake* (cluster: V-CMPL-V; cmpl: COMPLEMENTIZER) (DMC-AJ-A-32-00-00-0AA0-030A-A_15-00)

3. *Put the hydraulic systems in the* *safetied-for-maintenance* *configuration* (cluster: TVPAR-PREP-N) (DMC-AJ-A-32-00-00-01AAA-528A-A_022-00)

4. *General maintenance procedure (Simulation of the* *on-ground or in-flight* *configuration).* (cluster PREP-N) (DMC-AJ-A-32-00-00-03AAA-913A-A_023-00)

5. *The Main Gear System is a* *rearward-retractable* *landing gear installed in the two sponsons of the aircraft, left and right.* (cluster: ADVERB-ADJ) (DMC-AJ-A-32-11-00-0AA0-040A-A_020-00)

6. *There are two alternatives to do the procedure (weight-on-wheels or weight-off-wheels condition).* (cluster: N-PREP-TN) (DMC-AJ-A-32-21-71-03AAA-520A-A_020-00)

7. *Remove and discard the two O-rings.* (cluster: LETTER-TN) (DMC-AJ-A-32-21-71-01AAA-520A-A_020-00)

Another difficulty that we have had to cope with has been how to treat MPs in Simplified Technical English. This CNL recognizes as adjectives both –ing forms (VING) and –ed participles (VPAR) from verbs, as in *Lubricate the reamed hole,* where *reamed* is the past particle of the TV *ream* premodifying the TN *hole*. As a result, both types of deverbal adjectives, VING and VPAR, have been included as possible heads in MPs, which is an improvement with respect to Cortés-Rodríguez's proposal [2016a], in which they were not contemplated. Since both types of modifiers can also be formed out of TVs, we must also account for the variants, TVING and TVPAR, and include their corresponding AVMs within the GDE in ARTEMIS (Figure 6):

```
<Category Type="TVPAR">
<Attribute ID="Concept" />
<Attribute ID="Recip" />
<Attribute ID="Reflex" />
<Attribute ID="Template" />
</Category>
```

```
<Category Type="TVING">
 <Attribute ID="Concept" />
<Attribute ID="Recip" />
 <Attribute ID="Reflex" />
 <Attribute ID="Template" />
</Category>
```

Figure 6. AVMs for deverbal technical adjectives from in –ed (TVPAR) and –ing (TVING) in the GDE (ARTEMIS)

4. Referential phrase (RP) and Modifier Phrase (MP) rules for ASD-STE100

Section 5 presents the rules that are necessary for the effective parsing of ASD-STE100 and which incorporate all the idiosyncrasies that have been presented in Section 4 as regards the computational treatment of phrasal constituents in Simplified Technical English within ARTEMIS 2.0.

Let us start by presenting the compounding rule proposed above for CTNs, which captures the peculiarity of those instances of hyphenation processes that always give as output a compound technical name:

8. CTN →GRAM-N || GRAM-TN || GRAM-GRAM-N || GRAM-GRAM-TN

Since a noun cluster can consist of up to 3 words maximum (according to the first rule for noun clusters), and since hyphenation (second rule) produces one word, the most complex conceivable pattern that could be registered would be the following:

9. GRAM-GRAM-GRAM_{CTN} GRAM-GRAM-GRAM_{CTN} GRAM-GRAM-GRAM_{CTN}

In Figure 7 below, we show the analysis of a complex example of clustering (*main-gear-door retraction-winch handle,* Specification 2007: p. 1-2-3), which illustrates the pattern GRAM-GRAM-GRAM_{CTN} GRAM-GRAM_{CTN} GRAM_{TN} where the nucleus of the RP (the TN *handle*) is premodified by a complex MP whose nucleus is realized by a hyphenated 2-gram CTN (*retraction-winch*), which is at the same time premodified by a hyphenated 3-gram CTN (*main-gear-door*):

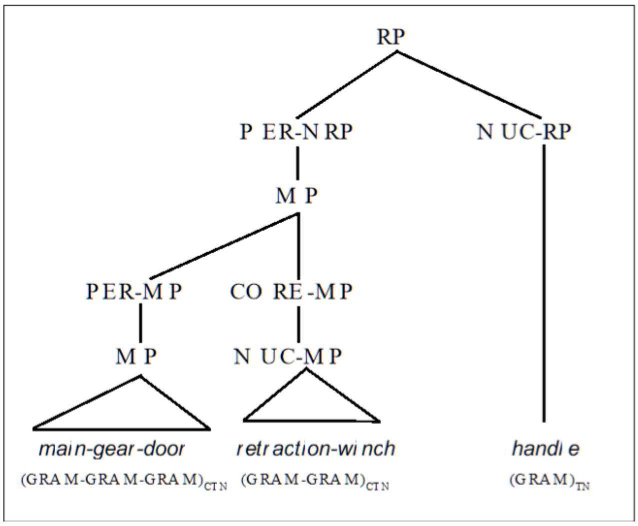


Figure 7. Analysis of a complex noun cluster

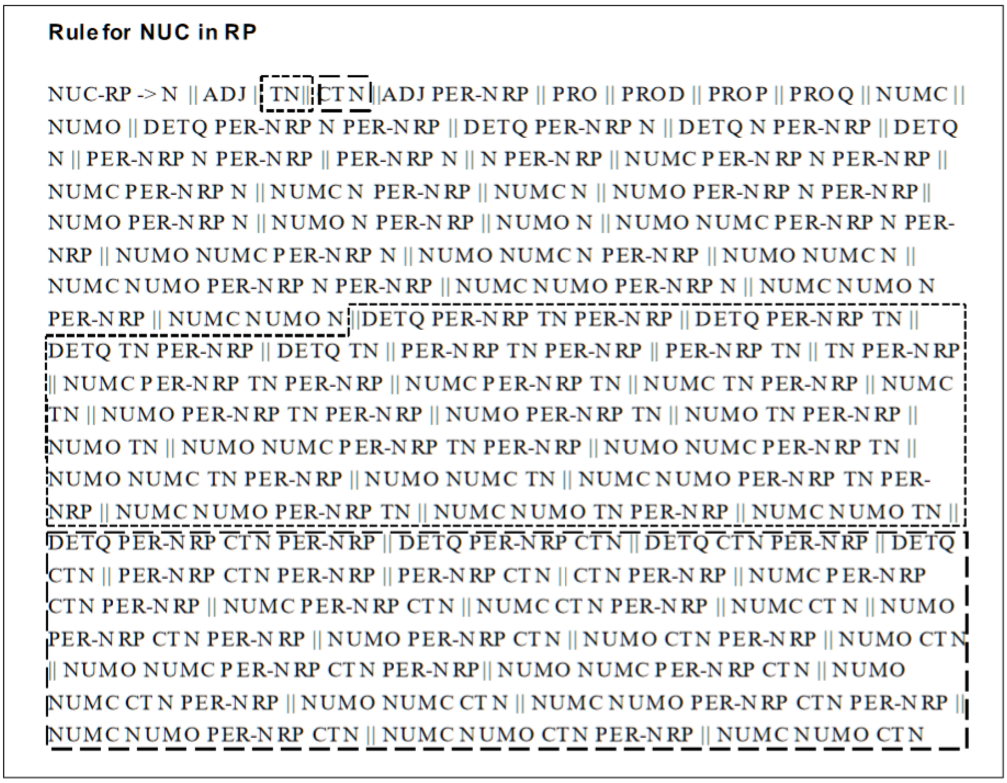
The introduction of the new POS TNs and CTNs, TVING and TVPAR, and the use of VING and VPAR as modifiers has led to a substantial modification of the rules proposed in Cortés-Rodríguez [2016a: 93-94] for the nucleus (NUC) of both RPs and MPs in ARTEMIS. We will first present the improved rules for the NUC of RPs that integrate the different configurations of RPs including either TNs or CTNs as NUC (inserted in boxes in the rule below (Figure 8) for better identification).

The following examples illustrate some of the possible syntactic patterns that are shown in square brackets:

10. ... *there are two KNEELING RESET push-button switches, [one*$_{PROQ}$*]*$_{NUC-RP}$ *in the Loadmaster Workstation (LMWS) and the [other*$_{PRO}$*]*$_{NUC-RP}$ *in the cockpit. (DMC-AJ-A-32-71-00-00AA0-040A-A_019-00)*

11. *Three*$_{NUMC}$ *cantilever*$_{PER-NRP(MP:TN)}$ *legs*$_{TN}$ *that operate independently*$_{PER-NRP(CL)}$ *[NUMC PER-NRP TN PER-NRP] (DMC-AJ-A-32-00-00-0AA0-030A-A_15-00)*

It is time now to present the improved rule for the NUC of MPs in ARTEMIS 2.0 that, as was the case in the rule for NUC in RPs, incorporates the new POSs that have had to be integrated to adapt these rules to the specificities of ASD-STE100, namely TN, CTN, VING, VPAR, TVING and TVPAR (Figure 9)[1].



Rule for NUC in RP

```
NUC-RP -> N || ADJ || TN || CTN ||ADJ PER-NRP || PRO || PROD || PROP || PROQ || NUMC ||
NUMO || DETQ PER-NRP N PER-NRP || DETQ PER-NRP N || DETQ N PER-NRP || DETQ
N || PER-NRP N PER-NRP || PER-NRP N || N PER-NRP || NUMC PER-NRP N PER-NRP ||
NUMC PER-NRP N || NUMC N PER-NRP || NUMC N || NUMO PER-NRP N PER-NRP||
NUMO PER-NRP N || NUMO N PER-NRP || NUMO N || NUMO NUMC PER-NRP N PER-
NRP || NUMO NUMC PER-NRP N || NUMO NUMC N PER-NRP || NUMO NUMC N ||
NUMC NUMO PER-NRP N PER-NRP || NUMC NUMO PER-NRP N || NUMC NUMO N
PER-NRP || NUMC NUMO N ||DETQ PER-NRP TN PER-NRP || DETQ PER-NRP TN ||
DETQ TN PER-NRP || DETQ TN || PER-NRP TN PER-NRP || PER-NRP TN || TN PER-NRP
|| NUMC PER-NRP TN PER-NRP || NUMC PER-NRP TN || NUMC TN PER-NRP || NUMC
TN || NUMO PER-NRP TN PER-NRP || NUMO PER-NRP TN || NUMO TN PER-NRP ||
NUMO TN || NUMO NUMC PER-NRP TN PER-NRP || NUMO NUMC PER-NRP TN ||
NUMO NUMC TN PER-NRP || NUMO NUMC TN || NUMC NUMO PER-NRP TN PER-
NRP || NUMC NUMO PER-NRP TN || NUMC NUMO TN PER-NRP || NUMC NUMO TN ||
DETQ PER-NRP CTN PER-NRP || DETQ PER-NRP CTN || DETQ CTN PER-NRP || DETQ
CTN || PER-NRP CTN PER-NRP || PER-NRP CTN || CTN PER-NRP || NUMC PER-NRP
CTN PER-NRP || NUMC PER-NRP CTN || NUMC CTN PER-NRP || NUMC CTN || NUMO
PER-NRP CTN PER-NRP || NUMO PER-NRP CTN || NUMO CTN PER-NRP || NUMO CTN
|| NUMO NUMC PER-NRP CTN PER-NRP|| NUMO NUMC PER-NRP CTN || NUMO
NUMC CTN PER-NRP || NUMO NUMC CTN || NUMC NUMO PER-NRP CTN PER-NRP ||
NUMC NUMO PER-NRP CTN || NUMC NUMO CTN PER-NRP || NUMC NUMO CTN
```

Figure 8. Syntactic rule for the NUC of RPs in Simplified Technical English in ARTEMIS 2.0

[1] The rest of the rules for the higher layers of RPs and MPs and those proposed for PPs in Cortés-Rodríguez [2016] remain the same.

NUC in MP

NUC-MP - > ADJ || ADV || N || RP || CL || S || TN || CTN || VING || VPAR || TVING || TVPAR

Figure 9. Syntactic rule for the NUC of MPs in Simplified Technical English in ARTEMIS 2.0

The following instances from the Airbus corpus include examples of the different types of NUC in MPs:

12. … the *kneeling*$_{\text{TVING}}$ actuators of each *MLG*$_{\text{TN}}$ *shock-absorber*$_{\text{CTN}}$ assembly (DMC-AJ-A-32-00-00-0AA0-030A-A_15-00)

13. A *sliding*$_{\text{VING}}$ tube assembly with the axle (DMC-AJ-A-32-00-00-0AA0-030A-A_15-00)

14. Do not let *compressed*$_{\text{VPAR}}$ gas touch your skin. (DMC-AJ-A-32-21-71-03AAA-520A-A_020-00)

15. …these sensors send the *unkneeled*$_{\text{TVPAR}}$ status signals when… (DMC-AJ-A-32-11-00-0AA0-040A-A_020-00)

Below we show the analysis of a complex RP (example (18)) that can be obtained by applying the parsing rules for RPs and MPs:

16. *a rearward-retractable landing gear installed in the two sponsons of the aircraft, left and right* (DMC-AJ-A-32-11-00-00AA0-040A-A_020-00.txt)

Figure 10 shows the first part of the analysis where we can observe how the head of the RP, which is the TN *landing gear* (NUC-RP), is, on the one hand, premodified at the level of the nuclear periphery (PER-NRP) by an MP whose nucleus is a two-word adjective, and, on the other hand, it is postmodified at the nuclear peripheral level (PER-NRP) by a restrictive MP whose nucleus is a VPAR (*installed*) that takes as argument (ARG-MP) a prepositional phrase (which is not further analysed for space restrictions).

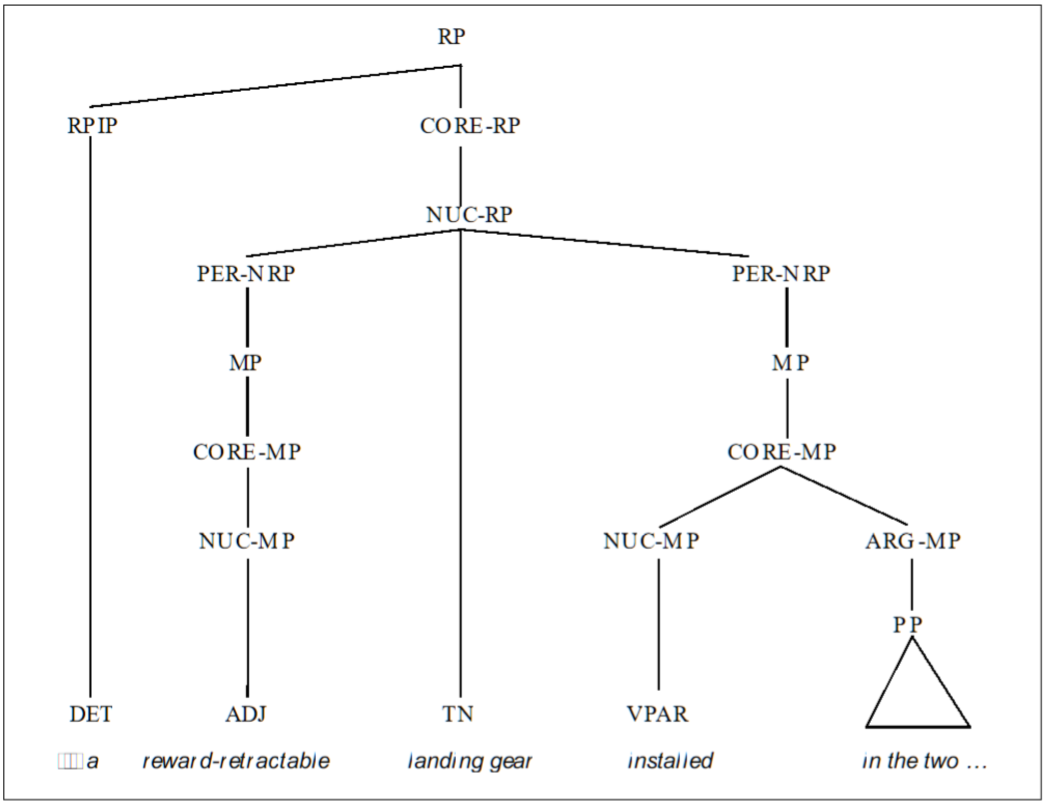


Figure 10. Analysis of a complex RP applying the parsing rules in ARTEMIS (I)

Note that RPs can indeed be very complex; in fact, we have found several cases of RPs that consist of more than 25 words, thus violating the maximum length allowed for whole sentences in the Specification document [Rule 6.3, p.1-6-4]. For instance, the following RP in example (19) consists of 27 words:

17. *The connections of the MLG unkneeled sensors to the RDC-1 and RDC-2, through which these sensors send the unkneeled status signals when the aircraft is not kneeled.* (DMC-AJ-A-32-11-00-0AA0-040A-A_020-00)

Conclusions

With this paper, we have tried to contribute to the development of the computational counterpart of the LCM by focusing on one of the components of the NLP parsing resource, ARTEMIS, namely, the Grammar Development Component (GDE), with the specific objective of advancing towards the computational treatment of phrasal constituents in a controlled natural language like ASD-STE100, Simplified Technical English.

With this aim in mind, we have tried to adjust the syntactic rules and AVMs for referential (RP) and modifier phrases (MP) that had already been presented in the GDE within ARTEMIS in Cortés-Rodríguez [2016a] to the specificities and requirements of Simplified Technical English. In particular, we have had to reflect in the GDE the fact that this Simplified English permits the use of technical vocabulary which is not included in the controlled dictionary of the specification document that lists the words that can be used in this type of clear, simple and unambiguous English. As a result, we have had to include new parts of speech (POS) with their corresponding attribute-value features so that this technical vocabulary could be analysed and recognised within the GDE: technical noun (TN), technical verb (TV), compound technical noun (CTN), and deverbal adjectives from technical verbs in –ing and –ed (TPAR and TVING respectively).

On the other hand, we have had to improve and adjust the parsing rules for phrasal realizations proposed in Cortés-Rodríguez [2016a] so that these rules could integrate the complex n-gram sequences that ASD-STE100 can produce as a result of complex and fully productive word-formation processes, such as the possibility of creating new words by using hyphenation or producing long noun clusters. The improved rules that have been introduced in this paper for the nucleus of RPs and MPs contemplate all these peculiarities and include the new types of POS.

With this research, we hope to have given ARTEMIS a validating platform in which to test this Simplified Technical English and also to have offered the users of ASD-STE100 a syntactic parser that is adapted to their needs.

References

ASD-STE Simplified Technical English. (2017). Specification ASD-STE 100. TM: International specification for the preparation of technical documentation in a controlled language. Issue 7. January 2017. Brussels: ASD.

Barrès, V. & Lee, J. (2014). Template construction grammar: from visual scene description to language comprehension and agrammatism. Neuroinformatics, 12(1), 181-208.

Bergen, B. & Chang, N. (2005). Embodied construction grammar in simulation-based language understanding. In J. Östman & M. Fried (Eds.), Construction grammars: Cognitive grounding and theoretical extensions, number 3 in Constructional Approaches to Language (pp. 147-190). Amsterdam / Philadelphia: John Benjamins.

Cortés-Rodríguez, F. (2016a). Towards the computational implementation of RRG. Círculo de lingüística Aplicada a la Comunicación, 65, 75-108.

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

*Díaz Galán, A. & Fumero Pérez, M. C. (*2017). ARTEMIS: State of the art and future horizons. In C. Rodríguez-Juárez (Ed.), Special Issue: New Insights into Meaning Construction and Knowledge Representation, Revista de Lengua para Fines Específicos, 23(2), 16-40.

Fumero Pérez, M. C. & 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.

FunGramKB. Functional Grammar Knowledge Base. URL: www.fungramkb.com.

Goldberg, A. (1995). Constructions: 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.

Kay, P. & Fillmore, J. (1999). Grammatical constructions and linguistic generalizations: The 'What's X doing Y?' construction. Language, 75, 1-33.

Kuhn, T. (2014). A survey and classification of controlled natural languages. Computational Linguistics, 40(1), 121-170.

Mairal-Usón, R. & Periñán-Pascual, C. (2009). The anatomy of the lexicon component within the framework of a conceptual knowledge base. Revista Española de Lingüística Aplicada, 22, 217-244.

Mairal-Usón, R. & Periñán-Pascual, C. (2016). Representing constructional schemata 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, F.J. (2009) Levels of description and explanation in meaning construction. In C. Butler & J. Martín Arista (Eds.), Deconstructing Constructions (pp. 135-198). Amsterdam / Philadelphia: John Benjamins.

Marques, T. & Beuls, K. (2016). Evaluation strategies for computational construction grammars. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Association for Computational Linguistics, pp. 1137-1146.

Martín Díaz, M. A. (2017) An account of English yes/no interrogative sentences within ARTEMIS. In C. Rodríguez-Juárez (Ed.), Special Issue: New Insights into Meaning Construction and Knowledge Representation, Revista de Lenguas para Fines Específicos, 23(2), 41-62.

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. (2007). Cognitive modules of an NLP knowledge base for language understanding. Procesamiento del Lenguaje Natural, 39, 197-204.

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

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. 165-196) 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.

Periñán-Pascual, C. & Mairal-Usón, R. (2011). The COHERENT Methodology in FunGramKB. Onomázein, 24, 13-33.

Ruiz de Mendoza Ibáñez, F. J. (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 grammar (pp. 231-270). Amsterdam / Philadelphia: John Benjamins.

Ruiz de Mendoza Ibáñez, F. J. & Galera Masegosa, A. (2014). Cognitive modeling. A linguistic perspective. Amsterdam / Philadelphia: John Benjamins.

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

Steels, L. (2004). Constructivist development of grounded construction grammars. In W. Daelemans (Ed.), Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (pp. 9-19). Barcelona. Association for Computational Linguistics.

Steels, L. (Ed.). (2011). Design patterns in Fluid Construction Grammar. In Constructional Approaches to Language, 11. Amsterdam / Philadelphia: John Benjamins.

Steels, L. & de Beule, J. (2006). A (very) brief introduction to fluid construction grammar. Proceedings of the 3rd Workshop on Scalable Natural Language Understanding (pp. 73-80). Association for Computational Linguistics. New York City.

Trott, S, Appriou, A., Feldman, J. & Janin, A. (2015). Natural language understanding and communication for multi-agent systems. In Artificial Intelligence for Human-Robot Interaction Papers from the AAAI Fall Symposium, (pp. 137-141).

Van Valin, R. D. (2005). Exploring the Syntax-Semantics Interface. Cambridge: Cambridge University Press.

Van Valin, R. D. (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. D. & LaPolla, R. (1997). Syntax. Structure, meaning and function. Cambridge: Cambridge University Press.

СИНТАКСИЧЕСКИЙ РАЗБОР СОСТАВЛЯЮЩИХ ФРАЗЫ В ASD-STE100 С ARTEMIS

Ф.Х. Кортес-Родригес[1], К. Родригес-Хуарес[2]
[1]Университет Ла-Лагуны (Тенерифе, Испания)
fcortes@ull.edu.es
[2]Университет Лас-Пальмас-де-Гран-Канария (Лас-Пальмас-де-Гран-Канария, Испания)
carolina.rodriguez@ulpgc.es

Данная работа сделана в рамках направления понимания естественного языка (ПЕЯ) с использованием устройства синтаксического анализа ARTEMIS, которое является прототипом ПЕЯ и состоит из трех суб-модулей: программа-конструктор (CLS) и программа-компоновщик (the COREL Scheme Builder) предоставляют семантические структуры, выделяя языковые фрагменты, программа развития грамматического окружения the Grammar Development Environment (GDE) занимается морфо-синтаксической компоновкой предложений. В ходе разработок были спроектированы правила порождения и условно-чистая матрица внутри GDE для анализа фразового компонента управляемого языка, ASD-STE100, а именно, упрощенного технического английского. Это полезно как для системы ARTEMIS, повышающей достоверность результатов, так и для пользователей ASD-STE100, т.к. программа становится более адаптированной к их нуждам.

Ключевые слова: *ARTEMIS, правила синтаксического анализа понимания естественного языка, упрощенный технический английский язык.*

Для цитирования: *Cortés-Rodríguez F.J., Rodríguez-Juárez C.* Parsing phrasal constituents in ASD-STE100 with ARTEMIS // Вопросы когнитивной лингвистики. 2018. № 3. С. 97-109.