# Modelling Grammatical and Lexical Knowledge: A Declarative Approach

Palmira Marrafa

University of Lisbon, Faculty of Arts
Department of Linguistics and CLG-Computation of Lexical
and Grammatical Knowledge Research Group (Centre of Linguistics)
Palmira.Marrafa@netcabo.pt

**Abstract.** This paper depicts the fundamentals of a computational grammar able to provide adequate representations for Portuguese simple sentences with several kinds of ambiguities. Besides the description of the architecture of the system proposed and the way it works, the paper focuses on the discussion of the nature of the specifications to encode in order to get a high level of precision. From a linguistic point of view, an endocentric phrase structure approach is adopted. The information is encoded in a DCG-like formalism, implemented in PROLOG.

## 1 Introduction

Modelling grammatical knowledge entails the specification of a large set of intertwined syntactic and semantic properties of linguistic expressions, which are highly structured and exhibit local and long distance dependencies ruled by several types of constraints.

In view of the complexity of the information to encode, the development of grammars that are suitable enough both for precision and coverage represents a great challenge.

As well-known, precision and coverage are conflicting requirements of natural language modelling, since a more precise grammar tends to be a more constrained grammar while constraints tend to reduce coverage (see [9] for a brief discussion of this trade-off).

Without neglecting coverage, this work is particularly concerned with precision, an essential requirement both for Theoretical Computational Linguistics central aims and for a wide range of applications.

Accordingly, the fragment of grammar presented here is able to rule out ill-formed expressions and inappropriate interpretations and to assign at least one representation to each well-formed expression for the constructions at issue. It covers the basic structure of simple sentences with several types of predication relations.

Such sentences frequently involve syntactic ambiguity, a major problem for computational natural language analysis.

Despite the complexity of the phenomena involved, the grammar has a suitable level of parsimony, since grammatical rules make appeal to the lexical entries which contain fine grained specifications of the syntactic and semantic restrictions imposed by the lexical units.

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 37 – 47, 2005.

The paper is organized as follows: Section 2 depicts the general architecture of the computational analysis system, the organization of the different modules it integrates and the way they interact; Section 3 is concerned with the analysis of the empirical data at stake and the kind of representations to be provided; Section 4 presents some results accounting for the descriptive adequacy of the system; finally, Section 5 presents some conclusion remarks.

## 2 System Architecture

Roughly, the system presented is a Definite Clause Grammar (DCG), implemented in PROLOG.

Similarly to what happens in most language technologies, the linguistic specifications and the computational procedures integrate different components. Let us observe Fig. 1, which provides a scheme of the overall system and the way the different components interact:

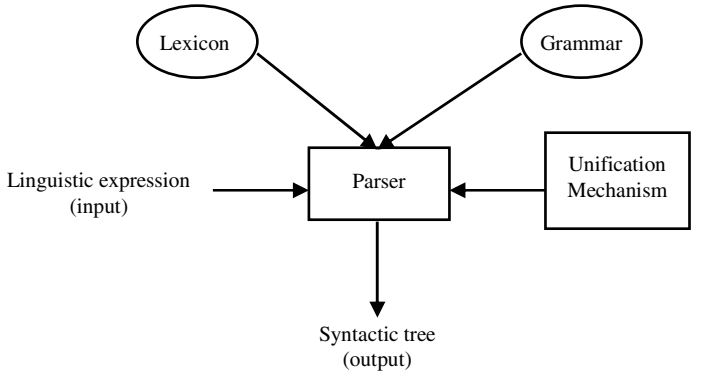


**Fig. 1.** General System Architecture

The lexical specifications and the grammatical rules are encoded in separated modules for the sake of an easier control.

The formalism used to encode the information is a DGC-like formalism with slight modifications. The main change concerns the fact that the right side of the rules is a list, in order to allow its interpretation by a bottom-up parser.

The option for a bottom-up parsing strategy has to do with the fact that several rules involve recursivity, namely the rules regarding modifiers, secondary predicates and complex predicates, as discussed in Section 3.

One of the components of the system is an unification mechanism, which deals with feature structures. It is worth to note that unification is a fundamental ingredient of the most used formalisms in Computational Linguistics because, among other reasons, it permits the easy encoding of information independently of any specific processing algorithm (on this matter, cf., for instance, [12]). The unification mechanism has a crucial role in this work, since it operates like a well-formedness filter, as illustrated in the next subsections.

### 2.1 Grammar Module

The syntactic configurations are defined in the Grammar module, which is basically a DCG, with slight modifications in order to allow its interpretation by a bottom-up parser. As referred to before, the main change concerns the fact that the right side of the rules is a list. The rules have the following format:

grammatical symbol ---> [any sequence of grammatical
symbols and control predicates].

Grammatical symbols are represented as follows:

<symbol designation> (<syntactic tree>, <feature structure>).

As mentioned, each lexical entry includes a feature structure that specifies the relevant properties of the corresponding lexical unit.

The control predicates used — *extract* and *unify* — guarantee the observation of the restrictions specified in the lexical entries. In other words, the control predicates guarantee that any linguistic expression whose syntactic structure is not consistent with the properties specified for the lexical items it integrates is ruled out. Both predicates are three-place predicates that make appeal to the feature structures included in the lexical entries: *extract(T,A,V)*; *unify(A,B,C)*.

Informally, *extract* is satisfied *if and only if T* is a feature that includes the attribute *A* and *A* has the value *V*; *unify* establishes that *A* and *B* unify in *C if and only if* for any attribute γ common to *A* and *B*, γ has the same value in *A* and *B*.

Let us examine two simple illustrative examples:

```
v1(v1(V,NP),T1)--->
   [v(V,T1),
    extract(T1,scat,[np]),
    np(NP,T2)].
```
(1)

```
np(np(Det,N),T)--->
   [det(Det,T1),
    n(N,T2),
    unify(T1,T2,T)].
```
(2)

In (1), the predicate *extract* guarantees that only verbs whose feature structure (*T1*) specifies the value *np* for the attribute *scat* enter in this rule. Therefore, it avoids the undesirable analysis of any *V NP* sequence as *V1*. It applies, for instance, to *joga este jogo* ("plays this game") but not to *joga esta manhã* ("plays this morning").

Through *unify*, the rule in (2) guarantees agreement between a noun and its determiner at the noun phrase level. The information specified in the feature structure of the latter (*T1*) and the information specified in the feature structure of the former (*T2*) have to unify. Consequently, anomalous expressions like *os gato* ('the_plural cat') are straightforwardly ruled out.

Since the syntactic structures are determined to a large extent by the properties of the lexical heads, specified in the Lexicon, the Grammar includes a relatively short number of rules. This way, the suitable economy of the system is preserved.

### 2.2 Lexicon Module

The descriptive adequacy of the representations provided by the system crucially depends on the information encoded in this module.

As rendered evident in Section 3, to capture the information to be included in the lexical entries is not a trivial task. A very fine grained syntactic and semantic characterization of the lexical units is required. Besides, the specifications for a given item have to be concerted with the specifications for many others.

Contrarily, encoding the information is relatively easy. Both syntactic and semantic properties are encoded by means of feature structures. Each property corresponds to an *attribute: value* pair. Let us exemplify with the abbreviated entries for *o* ('the_masc') and *dar* ("to give"):

```
det(det(o),[…,numb:sing,gend:masc,…])--->[o].        (3)
```

```
v(v(dar),[…,scat:[np,pp],…])--->[dar].               (4)
```

The determiner *o* is specified for the attributes *numb* (number) and *gend* (gender), whose values are *sing* (singular) and *masc* (masculine), respectively. The verb *dar* is specified for the attribute *scat* (subcategorization) whose value is the list *[np, pp]*.

Semantic properties can also be encoded in a similar way.

## 3 Linguistic Specifications

From a linear order point of view, the expressions treated so far in this project have the following skeleton: *N V N A*.

Despite the apparent simplicity of this sequence, it can correspond to several very distinct structures. Depending on several factors – in particular, the semantic properties of the elements involved –, the adjectival constituent can be interpreted as: (i) a secondary predicate oriented to the subject – (5a); a secondary predicate oriented to the object – (5b); a modifier within the NP object – (5c); part of a complex predicate – (5d). Let us observe some corresponding illustrative examples:

a. Ele dança a valsa descalço.
"He dances the waltz barefoot"

b. Ele viu a Maria furiosa. (5)
'He saw the Maria furious'
"He saw Maria furious"

c. Ele prefere o café brasileiro.
'He prefers the coffee Brazilian'
"He prefers Brazilian coffee"

d. Ele põe a Maria alegre.
'He makes the Maria happy'
"He makes Maria happy"

The syntactic and semantic properties of the structure corresponding to the different types of predications are discussed in the next sub-sections. The analysis is informed by

the assumption that syntactic expressions are endocentric. In other words, syntactic expressions are considered to be projections of lexical heads (so far, functional heads are not taken into account). The X-bar convention is adopted to represent syntactic configurations.

### 3.1 Secondary Predication

As illustrated above, secondary predication involves an argument of the primary predicate – the external argument in (5a); the internal argument in (5b) – and a non-verbal predicate (for the sake of simplicity of the explanation only APs are considered in this paper, but PPs and AdvPs can also be secondary predicates) which expresses an atomic event (a state, in other words) that occurs in the same temporal interval in which occurs the primary event. These circumstances justify the secondary predicates co-occurrence restrictions imposed by the head of the primary predicate, illustrated below:

a. Ele dançou descalço.
"He danced barefoot"
b. *Ele dançou arrependido.
"He danced regretful"
c. Ele chegou arrependido.
"He arrived regretful"

(6)

a. Ele convidou a Maria bêbeda
'He invited the Maria drunk'
"He invited Maria$_i$ drunk$_i$"
b. *Ele convidou a Maria indecisa.
'He invited the Maria undecided'
"He invited Maria$_i$ undecided$_i$"
c. Ele viu a Maria indecisa.
'He saw the Maria undecided'
"He saw Maria$_i$ undecided$_i$"

(7)

Marrafa [6] argues, along with the basic lines of Marrafa [3], that secondary predicates, lacking independent participants and time information, are a kind of "parasites" of primary predicates, as synthesized in (8), where *J* a temporal interval, *t1…ti…tn* sub-intervals of *J*, *e1…ei...en* the sub-events of a primary event that occur in *t1…ti…tn*, a secondary (atomic) event, *x* a participant in *e1…ei...en* and in $\varepsilon$, $p^k$ the set of properties assigned to *x* through *e1…ei...en* in *t1…ti…tn*, $p^m$ the set of properties assigned to *x* through $\varepsilon$ in *t1…ti…tn*.

In other terms, $p^m$, the set of properties associated to the event denoted by a secondary predicate (referred to above as secondary event), applies to a participant of the event denoted by a primary predicate in all the temporal sub-intervals in which its sub-events occur. That is, $p^m$ and $p^k$, the set of properties associated to *e1…ei...en*, apply to the same participant in the same temporal intervals. Consequently, $p^m$ has to be compatible with $p^k$. And the system has "to know" this.

$$\left[\begin{array}{l} J \\ e_1(x)^{pk}{}_{t1}\ldots ei(x)^{pk}{}_{ti}\ldots(en(x)^{pk}{}_{tn} \\ \varepsilon(x)^{pm}{}_{t1\cdots\, ti\cdots tn} \end{array}\right] \tag{8}$$

Let us now re-examine the example (5c), here renumbered (9a), in comparison with (9b):

(9)
a. Ele prefere o café brasileiro.
'He prefers the coffee Brazilian'
"He prefers Brazilian coffee / *He prefers the coffee Brazilian"
b. Ele prefere o café frio.
'He prefers the coffee cold'
"He prefers the coffee cold / He prefers the cold coffee"

As we can observe, (9b), but not (9a), is ambiguous between an interpretation where the adjectival constituent is a modifier of *café* ("coffee") and an interpretation where it is a secondary predicate oriented to the object. In (9a) the interpretation corresponding to *He prefers the coffee Brazilian* is excluded (in coherence with the ungrammaticality of the English expression). More precisely, in this case, the interpretation corresponding to the secondary predication is not available.

It seems obvious that the contrast above derives from the semantic properties of the adjectival constituents involved. Concretely, considering the dichotomy accidental properties *vs*. permanent or inherent properties (this distinction goes back to Milsark [7], [8] and Carlson [1]), the property denoted by *brasileiro* ("Brazilian") belongs to the latter class and the property denoted by *frio* ("cold") to the former one.

It is apparent from the data that secondary predication is only compatible with the expression of accidental properties. This restriction takes place also when the secondary predicate is oriented to the subject, as exemplified below:

(10)
a. Ele partiu feliz.
"He left happy"
b. *Ele partiu alto.
"He left tall"

However, the characterization of the adjectives on the basis of this dichotomy is not straightforward, since adjectives can be ambiguous in relation to those properties, as it is the case of *triste* ("sad") in the examples below:

(11)
a. Ele encontrou a rapariga triste.
"$He_i$ met the girl $sad_i$"/ "He met the sad girl"/ "He met the $girl_i$ $sad_i$"
b. Ele leu o livro triste.
"$He_i$ read the book $sad_i$"/ "He read the sad book"/ "*He read the $book_i$ $sad_i$"

Both sentences are ambiguous, but (11a) has one interpretation more than (11b). In the former sentence *triste* can be secondary predicate of *Ele* ("He") and both modifier and secondary predicate of *rapariga* ("girl"), while in the latter one the interpretation of secondary predicate oriented to the object is not available.

Despite their complexity, all the restrictions have to be encoded in order to avoid both over- and under-generation of representations.

Regarding the syntax of these constructions, the co-occurrence restrictions imposed by the verb to the secondary predicates suggest that they are not excluded of the maximal projection of *V*.

Nevertheless, the restrictions imposed to the predicates oriented to the object are stronger than those imposed to the predicates oriented to the subject. Moreover, the order *predicate oriented to the object* < *predicate oriented to the subject* is obligatory, as rendered evident by the contrast below:

a. Ele bebeu o café frio triste.
'He drank the coffee cold sad'
b. *Ele bebeu o café triste frio.
'He drank the coffee sad cold' (12)

On the basis on these facts, the predicates oriented to the object are represented in adjunction to *V1* (the numeric notation is used for the sake of coherence with the notation used in the modelling formalism) and the predicates oriented to the subject in adjunction to *V2* (*VP*), as shown in (13):

$$[\ldots[_{V2}[_{V2}[_{V1}[_{V1}[_{V}]\ldots][_{pred\_object}]]][_{pred\_subject}]]] \quad (13)$$

It is worthwhile to note that this representation also satisfies the subject-predicate reciprocal m-command constraint extensively argued for by Marrafa [3] and further related work, but the discussion of this issue is not within the goals of this paper.

### 3.2 Complex Predicates

Concerning complex predicates, this paper focuses on lexical-conceptual structure deficitary verbs and follows mainly Marrafa's [4] and [5] proposals and previous related work.

In order to clarify the concept of lexical-conceptual structure deficitary verb let us start by examining the following example:

Ele pintou a parede de amarelo.
"He painted the wall yellow" (14)

The situation described in (13) entails that *a parede* ("the wall") became *amarela* ("yellow") as a result of painting. This means that the verb denotes an event with a definite endpoint (cf. Wechsler [13], among others). In other terms, the verb denotes a transition event (in the sense of Pustejovsky [10], [11]), which is structured as stated below:

$[_T [_P\ e_1 ...e_n]\ e_m]$
T, Transition; P, Process; e, atomic event; $e_m > e_n$; $e_m \neq e_1$ (15)

Accordingly, the sentence in (14) has the following lexical-conceptual structure (Pustejovsky's LCS'):

LCS' { [[act(ele,parede)&~pintada_de_amarelo(parede)],
[pintada_de_amarelo(parede)]]
"{ [[act(he,wall)&~painted_yellow(wall)],
[painted_yellow(wall)]]" (16)

As it becomes evident, the verb plus the resultative expression, *de amarelo* ("yellow"), form a lexical-conceptual unit, that is, a complex predicate, as extensively argued by Marrafa [3].

The absence of the resultative does not have any impact on the LCS', as we can observe:

a. Ele pintou a parede.
"He painted the wall"

b. LCS' { [[act(ele,parede)&~pintada(parede)],
[pintada(parede)]]
"{ [[act(he,wall)&~painted(wall)],
[painted(wall)]]" (17)

Let us now consider again the example in (5d) (here renumbered as (18a)):

a. Ele põe a Maria alegre.
'He makes the Maria happy'
"He makes Maria happy" (18)

The LCS' associated to it seems to be (18b) and not (18c).

b. LCS' { [[act(ele,Maria)&~feliz(Maria)],
[feliz(Maria)]]
"{ [[act(he,Maria)&~happy(Maria)],
[happy(Maria)]]"

c. LCS' { [[act(ele,Maria)&~tornada_feliz(Maria)],
[tornada_feliz(Maria)]]
"{ [[act(he,Maria)&~made_happy(Maria)],
[made_happy(Maria)]]"

This suggests that *Q* is instantiated just with the resultative. It is then expected that the absence of the resultative induces ungrammaticality, in coherence with the facts:

*Ele põe a Maria.
'He makes the Maria' (19)
"He makes Maria"

Along the same basic lines of Marrafa [3] and further work, verbs like *pôr* ("make") are argued here to be LCS' deficitary, in the sense that they do not include in their denotation the set of content properties of the final state of their LCS', as stated below:

Informal definition:

$\forall$v((verb(v), $\exists \mathcal{E}$ , LCS'_of_v( $\mathcal{E}$ ), $\exists$e, final_state(e), (20)
e $\subset \mathcal{E}$ , $\exists$ $\pi$, set_of_semantic_features_of($\pi$,e), $\pi = \varnothing$)
=> LCS'_deficitary(v))

Since that set is empty, the LCS' cannot hold an appropriate interpretation. A syntactic structure that projects an anomalous LCS' is, then, previewed to be ruled out (it does not satisfy the requirement of full interpretation).

In this case, the resultative fills the gap of the LCS' of the verb (cf. the contrast between (18a) and (19)).

Therefore, these facts show that the representation of the predicates at issue has to include information concerning the resultative expression.

Regarding the syntactic structure, the general internal structure of *V2* (*VP*) is largely inspired in Larson's [2] proposal. In what specifically concerns the complex predicate, it has to be represented as a syntactic unit to account for the data discussed above. That means that both elements have to be immediately dominated by the same node. Therefore, the non-verbal part of the predicate is represented in adjunction to *V*. To derive the canonical order, *V* moves to a higher position. This movement is captured by co-indexation between *V* and the *t*(race) that occupies its basic position. The internal structure of *V2* for complex predicates is, then, the following:

$[_{V2} \ldots [_{V1} V_i \ [_{V2} \ NP \ [_{V1} \ [_{V} [_{V} t_i] \ AP]]]]]$ (21)

A major problem that the computational analysis system has to deal with is the high level of ambiguity induced by certain polysemous verb forms with resultative interpretations, as it is the case of *deixar* ("to let"), referred to in the next section.

## 4 Results

The representations provided by the system for two ambiguous sentences are presented below. Constituents in a subject-predicate relation, a head-modifier relation or belonging to a complex predicate are marked in the syntactic tree by means of *co I*, *I* instantiated with the same value.

In spite of their superficial similarity, these two sentences have not the same level of ambiguity. In the case of (22) four interpretations are available, while in the case of (23) there are only two.

O Jorge deixou a rapariga triste.
'the Jorge left the girl sad'
"Jorge left the girl sad"
Interpretations: (i) deixou triste ≡ entristeceu ("made sad"): [*deixou triste*] complex predicate (22)
(ii) [$_{NP}$a rapariga triste]: ap modifier
(iii) [$_{NP}$ a rapariga$_i$][triste$_i$]: ap pred-obj
(iv) [$_{NP}$ o Jorge$_i$][triste$_i$]: ap pred-subj

O Jorge leu o livro triste.
'the Jorge read the book sad'
"Jorge read the book sad" (23)
Interpretations: (i) [$_{NP}$o livro triste]: ap modifier
(ii) [$_{NP}$ o Jorge$_i$][triste$_i$]: ap pred-subj

([*deixou triste*] complex predicate)

```
f(f(np(det(o),n(jorge)),
    v2(v1(v(deixou)co_147,
    v2(np(det(a),n(rapariga)),
    v1(v(v(t)co_147,
    ap(a(triste)))))))));
```

([$_{NP}$a rapariga triste]: ap modifier)

```
f(f(np(det(o),n(jorge)),
    v2(v1(v(deixou),
    np(det(a),n(rapariga),ap(a(triste)))co_178)
                                        co_178)));
```

([$_{NP}$ a rapariga$_i$][triste$_i$]: ap pred-obj)

```
f(f(np(det(o),n(jorge)),
    v2(v1(v1(v(deixou),
    np(det(a),n(rapariga))co_178),
    ap(a(triste))co_178))));
```

([$_{NP}$ o Jorge$_i$][triste$_i$]: ap pred-subj)

```
f(f(np(det(o),n(jorge))co_78,
    v2(v2(v1(v(deixou),
    np(det(a),n(rapariga)))),
    ap(a(triste))co_78))).
```

([$_{NP}$o livro triste]: ap modifier)

```
f(f(np(det(o),n(jorge)),
    v2(v1(v(leu),
    np(det(o),n(livro),ap(a(triste)))co_174)
                                    co_174)));
```

([$_{NP}$ o Jorge$_i$][triste$_i$]: ap pred-subj)

```
f(f(np(det(o),n(jorge))co_75,
    v2(v2(v1(v(leu),
    np(det(o),n(livro)))),
    ap(a(triste))co_75))).
```

As we can observe, the representations provided are the adequate ones. All the licensed interpretations, and only the licensed interpretations, are assigned a representation.

## 5 Conclusions

Despite the intricacy of the syntactic and semantic restrictions of the constructions at stake, the system presented here is able to provide adequate representations, accounting for structural ambiguity.

The use of feature structures to specify lexical information allows to encode the information related to such restrictions in a very fine grained way.

In view of the richness of lexical entries, interfacing lexical descriptions with grammar rules allows for a relatively parsimonious grammar.

The modular and declarative formulation adopted greatly facilitates the extension of the grammar to another kind of structures, as well as porting it from one formalism to another.

## References

1. Carlson, G.: Reference to Kinds in English. PhD dissertation, University of Massachusetts-Amherst (1977).
2. Larson, R.: On the Double Object Construction. Linguistic Inquiry, 19, (1988) 335-391.
3. Marrafa, P.: Computação de Ambiguidades Sintácticas: evidências em favor dos modelos baseados em conhecimento linguístico. IN COGNITO (forthcoming).
4. Marrafa, P.: Extending WordNets to Implicit Information. Proceedings of LREC 2004, International Conference on Language Resources and Evaluation. Lisboa, Portugal, (2004) 1135-1138.
5. Marrafa, P.: The representation of Telic Complex Predicates in Wordnets: the case of lexical-conceptual structure deficitary verbs. Proceedings of Convergences03, International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies. Library of Alexandria, Alexandria, Egypt (2003).
6. Marrafa, P.: Predicação Secundária e Predicados Complexos em Português: Análise e Modelização. PhD dissertation. University of Lisbon (1993).
7. Milsark, G.: Existential Sentences in English. PhD dissertation, MIT (1974).
8. Milsark, G.: Toward an Explanation of Certain Pecularities of the Existencial Construction in English. Linguistic Analysis, 3, (1977)1-29.
9. Pereira, F.: Sentence Modeling and Parsing. In G. Varile & A. Zampolli (eds) Survey of the State of the Art in Human Language Technology, 130-140. National Science Foundation and European Comission (1995).
10. Pustejovsky, J.: The Syntax of Event Structure. Cognition, 41, (1991) 47-81.
11. Pustejovsky, J.: The Generative Lexicon. Cambridge, MIT Press, Mass. (1995).
12. Uszkoreit, H. & Zaenen, A.: Language Analysis and Understanding. In G. Varile & A. Zampolli (eds) Survey of the State of the Art in Human Language Technology. National Science Foundation and European Commission (1995) 130-140.
13. Wechsler, S. An Analysis of English Resultatives Under the Event-Argument Homomorphism Model of Telicity. Proceedings of the 3rd Workshop on Text Structure. University of Texas (2000).