UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

–accusative and ablative–, we have preferred to add dative.

When we refer to ablative, it is a governed ablative. It is not what we call adverbial complement in traditional grammar. We attach a letter to each case. Accusative is q, dative is t and ablative is r. A non-argumental complement will be named o_L. We will just consider these four argumental relations, plus $\odot$, which means absence of argument. All together they form set ℓ, which will have four elements and is the set of arguments that can afford focus v, this is to say, the levels that it can assign. $\ell = \{q, t, r, \odot\}$. Starting from ℓ a new set is defined, for focus o, which is the set of levels that it can take on, $^{-}\ell = \{q, t, r, L\}$.

We observe now the set ℓ. $|\ell|$ is the number of elements of this set, 4. $\ell(v)$ is the subset ℓ for any v. $|\ell(v)|$ is the number of level assignments of $\ell(v)$.

In the same way $|^{-}\ell|$ is the number of elements of the set $^{-}\ell$. $|^{-}\ell| = 4$. $^{-}\ell(o)$ is the subset of $^{-}\ell$ for o. $|^{-}\ell(o)|$ is the number of level assignments of $\ell(o)$. Every o can take on more than one level.

One v can bring about many o on its right side, because we have defined an o_L as that one that can be freely inserted on the right of o without being at all necessary from a syntactic point of view, but neither rejected. We do not consider o_L elements in level assignments relations. We symbolize valency v with a superscript. So, v^0 is a v with valency 0, which means that it cannot accept anything on its right. v^1 denotes a v that only wants an o on its right, and v^2 is a v that wants two foci o on its right. Most commonly, the valency on the right of v is 1. And the existence of v like $v^{0,1}$ –or any other valencies combination– is even more usual. We denote levels as subscripts. So, that is how the matter stands, v^1_q, o $v^2_{(T,r)}$, although due to simplicity reasons, superscripts may be removed when the set of all subscripts is stated.

$|\ell|$ set and valency are not equivalent. If the valency is 0, $|\ell| \neq 0$, $|\ell| = 1$. Valency 0 is denoted by $\odot$. Furthermore, let's imagine that one $v^{1,2}$ exists where v^1_q and $v^2_{s,t}$. Then, the set of elements in ℓ would be 3, whereas the

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

number of valencies would be 2, and the value of every valency would be 1 and 2. The $\ell(v)$ set contains all the subscripts that v can take when it operates with 0, 1 or 2 valencies.

Elements of ℓ are very difficult to be specified. Who selects the level that has to be attributed to one o? We imagine the process in the following way. We have a new created string with the following elements:

$$sv_{(\odot,r)}o_{(r,L)}$$

Since $\ell(v) \cap^{-} \ell(o) = r$, then the element that v will impose on o will be r, because it is the common element. So, they form the string $sv_r o_r$. This one could be the case of a sentence such as *Viu a Barcelona.*

On the other hand, let's imagine this same focus o with one $v_{(\odot,q)}$, in the following string:

$$sv_{(\odot,q)}o_{(r,L)}$$

Since $\ell(v) \cap^{-} \ell(o) = \varnothing$, then v assigns $\odot$ and o takes L, forming the sentence $sv_{\odot}o_L$ which may denote, for example: *Ells canten a Barcelona.*

5.3.4 Kinds of linguistic strings

We will take two criteria into account in order to distinguish different types of strings:

- Their structure.
- Their condition regarding application of the rules.

5.3.4.1 Classification of strings according to the form of their elements

The definition of strings must hold more features than the definition of patterns, because the relations of *agreement* and *level assignment* take place in

ISBN:978-84-691-1896-2/D.L:T-352-2008

the stratum string, and not in the stratum variables. All of them are defined starting from the *basic string*. It means that if we do not affirm the contrary, they follow the same rules of *agreement* and *level assignment*. Every linguistic string x is composed by the following vocabulary $V_x = \{s, v, o\}$. Taking into account what we have just said, we can define:

- **Basic string**: it is that one accomplished by means of replacing every variable of a basic pattern by a focus belonging to its domain. The next formation criteria follows:
  - (i) Simplicity: it has only one occurrence for each element.
  - (ii) Precedence: svo is an ordered string. For a SVO language such as Catalan:
    - $s \prec v, o$
    - $v \prec o$
  - (iii) Agreement: $PN_s = PN_v$
  - (iv) Level assignment: $\exists l \; |l \in \ell v, l \in \, ^{-}\ell o$
- **Reverse string** is that one in which:
  - (i) Formation criteria are the same as in the basic string, except:
  - (ii) precedence criterion is not respected.
- **Compound string** is that one in which:
  - (i) $\exists u \mid |u| > 1$
  - (ii) precedence criterium is not taken into account.
- **Poli-v string** is that one in which:
  - (i) $|v| > 1$

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

- **Partial tring** is that one in which:
  (i) $0 \leq |s| \leq 1$
  (ii) $0 \leq |v| \leq 1$
  (iii) $0 \leq |o| \leq 1$
  (iv) $1 \leq |x| \leq 2$
- **Minimal string** is a kind of partial string in which:
  (i) $len(x) = 1$
- **Minimal in v string** is a kind of minimal string in which:
  (i) $\exists v | v \in x$
- **Compound partial string** is that one in which:
  (i) $len(x) \geq 2$
  (ii) $\exists u| \; |u| = 0$
  (iii) $\exists u'| \; |u'| \geq 2$
- **Compound minimal string** is a kind of compound partial string in which:
  (i) $\exists u| \; |u| = 0$
  (ii) $len(x) \geq 3$
- **Complex string** is that one which differs from a basic string in the simplicity criterion.

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

5.3.4.2 Classification of strings according to their state with respect to the application rules

Strings may be classified following these criteria: (a) whether they are the result of applying a rule or not, and (b) in such case, which is their position in the derivation. According to this, we want to distinguish four kinds of strings:

a) **Primary string**: is that one in which no rule has been applied. We establish that every primary string is, by definition, a basic string in all the operations we will carry out in this thesis.

b) **Derivative or Resultant string**: is the outcome of some operation.

c) **Stage string**: is that one in which a rule has been applied in order to obtain a new string.

d) **Closing string**: is the last string in a derivation. Taking into account the kind of operations applied, different criteria will have to be established to validate the closing strings.

5.3.5 Results

In this section we have defined a string and made a classification. If, as it seems, at first the structure of strings depends on patterns –leaving aside the relations among the foci which do not exist at variables level– we can affirm that this same classification is also suitable for them. It is to say, we can speak about minimal pattern in the same way we can speak about minimal string. This terminology will be applied to strings in almost exclusive measure along our syntactic proposal, but nothing prevents it from sometimes applying to patterns. On the other hand, we have seen that strings have very strong internal formation conditions among the focus that compound them. It means that the string, as a final production, is a much more complicated

ISBN:978-84-691-1896-2/D.L:T-352-2008

structure than the pattern and, hence, it needs more working out. Related to the suggested classification and the crossing of both, we can deduce some interesting aspects:

1. Any primary string is, by definition, a basic string.

2. Any compound string is derivative.

3. Any minimal string is derivative.

4. A closing or derivative string cannot be primary, but it can be basic.

5.4 ULPS

Let's consider a sentence formed from a basic pattern which is unfolded in:

- the pattern with the variables S (subject), V (Verb), O (Object),
- the terminal string of linial production.

The graphical representation of this structure with two strata can be the following:

s	v	o
S	V	O

Figure 5.2: ULPS

In a formal way:

$$\begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix}$$

This representation, in which we take into account two strata of the same sentence, leads to what we call *unfolded linguistic pattern structure* (ULPS). We use the Greek alphabet in capital letters to designate the ULPS – $X, \Upsilon, Z, \ldots$

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

The representation of each variable with a focus of its domain is called *pole*.

$$\begin{array}{c} v \\ \boxed{\;\;\; V \;\;\;} \end{array}$$

Figure 5.3: Pole of a ULPS

In a ULPS that represents a basic pattern we find, then, three poles:

$$\begin{pmatrix} s \\ S \end{pmatrix}, \begin{pmatrix} v \\ V \end{pmatrix}, \begin{pmatrix} o \\ O \end{pmatrix}$$

Later on, when we carry out links with staggered cuts, the existence of other poles with their combination rules will be tested.

5.4.1 Axiomatic ULPS in Catalan

We call axiomatic ULPS of a language all those which exist without any process of linguistic recombination. Making use of the terminology we have used for the string, axiomatic ULPS would be equivalent to the primary ones. We have considered that the axiomatic pattern in Catalan is made up of SVO, and therefore, the first evident *axiomatic ULPS* is the *basic* one. Nevertheless, we can suppose the existence of some axiomatic structures which are even smaller.

5.4.1.1 ULPS of shortened S

Let's consider sentences such as:

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

Example 1

Hi ha molts gats.

`There are many cats.`

In English, we have a sentence with subject placed on the right. In Catalan, this structure is considered impersonal. The phrase *molts gats* (`many cats`) occupies the place and function o within the sentence. We could simply think that we have $\begin{pmatrix} v & o \\ V & O \end{pmatrix}$, this is to say, a structure which does not have the variable S and with a v without subscripts PN. Eventhough, *hi ha molta gent* (`there are many people`) does not correspond to $\begin{pmatrix} v & o \\ V & O \end{pmatrix}$ structure, that is, what we call *infinitive*, because we would need the form *haver-hi molta gent* (`there be many people`), which is impossible in a simple sentence in Catalan. So, the string that represents this sentence in Catalan is the following: $\lambda_{3e}v_{3e}o$. This is a case in which the string $\lambda \in S$ is positioned as a subject, so s exists. However, it seems obvious that the subject position in Catalan, in this case, is actually empty.

There is no other way, then, to reach the conclusion that in Catalan the existence of subject s in simple sentences is essential, despite the lack of the variable. The fact that this λ cannot be occupied under any circumstance by another focal string, leads us to believe that this is the case of a λ which only exists because it is compulsory to accomplish the agreement criterion that is essential in the formation of a string. We can consider an s that does not arise from any variable within the pattern. It is a ULPS with a staggered end, with the shortened $\begin{pmatrix} s \\ S \end{pmatrix}$ pole, denoted by $\begin{pmatrix} s \\ \end{pmatrix}$. We call this structure *ULPS of shortened S*

The fact of assuming this hypothesis involves the acceptance in Catalan of the existence of primary structures with overhangs, considering that no single operation has been applied.

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

s	v	o
	V	O

Figure 5.4: ULPS of shortened S

5.4.1.2 Partial ULPS

Let's examine now another sentence:

Example 2

En Joan va morir.

`John died.`

In this example it seems clear that, in English as well as in Catalan, there does not exist any focus o. Apparently subscripted λ does not exist in $\mathcal{O}$, because there are some v that do not need anything on its right, such is the case for the verb *morir* (`to die`). Let's launch into the hypothesis that there is a kind of structures which are formed by strings $v_\odot$ that have neither variable O nor focus o. This is the easiest explanation for this kind of sentences. So, some axiomatic structures in Catalan with a missing pole $\begin{pmatrix} o \\ O \end{pmatrix}$ are accepted, being fully aware that no cut has been applied to them. We call these structures *Partial ULPS*.

s	v
S	V

Figure 5.5: Partial ULPS

5.4.1.3 Partial shortened ULPS

Finally, let's consider another sentence:

Example 3

Plou.

`It rains.`

This sentence shows the two phenomena we have just seen. On the one hand, it shows the $\begin{pmatrix} s \\ \ \end{pmatrix}$ shortened pole; and on the other hand, it shows the lack of $\begin{pmatrix} o \\ O \end{pmatrix}$. Therefore, it seems that primary structures with this form exist in Catalan:

$$\begin{pmatrix} s & v \\ & V \end{pmatrix}:$$

These structures are called *Partial shortned ULPS*

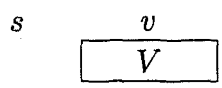


Figure 5.6: Partial shortned ULPS

5.4.2 Results

In this section a ULPS has been defined, which is only a representation of an unfolded pattern. We have also established what is a pole, and we have seen that a basic ULPS is three-poled. ULPS is the most complex linguistic structure because it has the formation requirements of patterns and strings. Moreover, we have defined four kinds of axiomatic ULPS in Catalan:

- basic ULPS,
- ULPS of shortened S,

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

- partial ULPS,
- partial shortened ULPS.

We do not want to simplify by establishing a relation for each ULPS with different strings. For example, it is obvious that the basic ULPS does not only correspond to a basic pattern, but also to a basic string. However, the shortened ULPS corresponds to a basic string, but not to a basic pattern. We need to take this into account when identifying a basic string with a shortened ULPS. On the other hand, a partial ULPS corresponds to a partial string but not to all partial strings. Let us we review the definition of this one. Even though it corresponds to a sv string, it should be with special features. For example, $\odot \in \ell$. On the other hand, a string of a shortened ULPS corresponds also to a partial string, but its pattern is minimal in v.

We defend that all these ULPS are axiomatic, and although their existence is foreseen, they are unusual structures which cannot act as a base for the study of general phenomena of language and, in any case, they deserve separated attention. When establishing which base we will choose for the operations we want to apply to different linguistic structures, we refer to the established condition in 5.3. If every primary string is basic by definition, then every primary ULPS will also be basic by definition. Appearance conditions of other ULPS that may not be basic will be established later if it is necessary. On the other hand, further on we will find some ULPS like these we have seen here, but obtained by means of cuts in basic strings. The most important difference between the former and the latter is that the ones we are dealing with at the moment are axiomatic and primary strings, and those we will see later are derivative. Whereas axiomatic strings correspond to structural facts of the language, a string like the one we have just seen results from a derivation, it would hardly bear the conditions that must be impossed to a closing string in order to be acceptable.

At this time, it seems obvious that when we refer to these ULPS, that, we think exist in Catalan, we will only shelter the basic one due to a methodological restriction reason.

s	v	o
S	V	O

Figure 5.7: Basic ULPS

5.5 Conclusions

At the end of this chapter we have achieved the minimal unit of our study: the basic ULPS. We have described it as a three-pole unit, each one of its elements, arranged by means of a rule of precedence, consists of a variable and a focus. The joint of foci of a ULPS which are catalyzed by rules of internal structure is called a string, and the joint of variables is called a pattern.

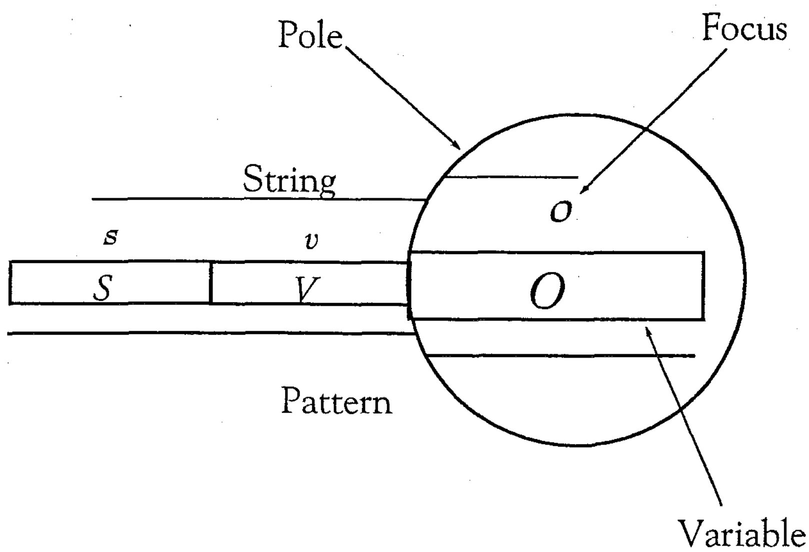


Figure 5.8: Analysis of the components of a ULPS

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

Chapter 6

Bases of syntactic rules: cuts and links

6.1 Cuts and syntactic pieces

6.1.1 Objectives

Many molecular computation systems are based on cutting and pasting. If we want to apply them to Linguistics, we must also carry out these two operations. We have established that the minimal unit of our work will be a *basic ULPS* and we have defined it. Now we try:

- To carry out as many cuts as possible in a ULPS.
- To establish which are the *syntactic pieces* caused by cuts, upon which we will carry out the joining operations in this thesis.

Taking as a model what happens inside a DNA molecule by the action of an enzyme, we will carry out two kinds of cuts in the basic ULPS:

- Blunt: the one that cuts whole poles.

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

- Staggered: the kind of cut that cuts short at least one pole by separating the variable from the focus.

At this point, we want to distinguish between *simple cuts* and *complex cuts.* A simple cut is applied to a string if it is cut only one time; a complex cut is applied to a string if it is cut two or more times. If we do not mention the opposite, only simple cuts will be applied in this chapter.

This distinction can also be applied to links. A *simple link* is one in which each of the strings that takes part in it is only joined by one side. A *complex link* is one in which at least one of the strings that takes part in it is joined by both ends.

Starting from these concepts, we can introduce two new ones: *simple operation* and *complex operation.* A simple operation is one in which each of the strings that take part is only subjected to one cut and one link. A complex operation is one in which at least one of the strings that takes part is subjected to more than one cut or more than one link.

Simple cuts will result in syntactic bits. All those pieces that contain at least one variable and at least one focus will be considered *syntactic pieces.*

Once we have obtained an inventory of syntactic pieces, we will consider *basic syntactic pieces* to be those obtained by means of simple cuts which can be combined; this is to say, when the fact of joining them with other pieces causes, at least, one correct result.

6.1.2 Blunt cuts

In a basic ULPS a blunt cut (shown by #) can be carried out in two positions. We obtain two pieces called "p" in each one of the cuts:

a) Cut $\begin{pmatrix} v & \# & o \\ V & \# & O \end{pmatrix}$

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

s	v	o
S	V	O

s	v	|	o
S	V	|	O

p:

s	v
S	V

p:

o
O

The formal representation is :

$$\begin{pmatrix} s & v & \# & o \\ S & V & \# & O \end{pmatrix} \Longrightarrow \begin{pmatrix} s & v \\ S & V \end{pmatrix}, \begin{pmatrix} o \\ O \end{pmatrix}$$

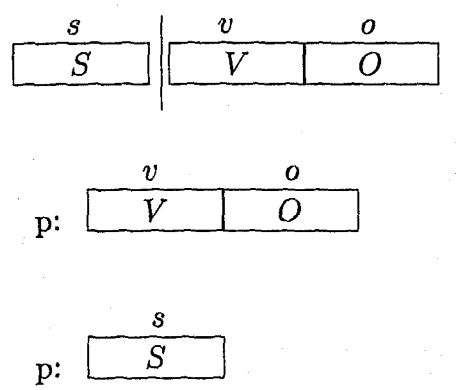
When the blunt cut breaks the string and the variables in the same place, then, for the sake of simplicity, we will adopt a notation where we will only take the strings into account. Thus, the most common way to represent this cut will be:

$$\begin{array}{l} s \;\; v \;\; \# \;\; o \longrightarrow s \;\; v \\ s \;\; v \;\; \# \;\; o \longrightarrow' o \end{array}$$

b) Cut $\begin{pmatrix} s & \# & v \\ S & \# & V \end{pmatrix}$

s	v	o
S	V	O

ISBN:978-84-691-1896-2/D.L:T-352-2008



Formal representation:

$$\begin{pmatrix} s & \# & v & o \\ S & \# & V & O \end{pmatrix} \Longrightarrow \begin{pmatrix} s \\ S \end{pmatrix}, \begin{pmatrix} v & o \\ V & O \end{pmatrix}$$

With the usual notation by blunt cuts:

$$\begin{array}{llllll} s & \# & v & o & \longrightarrow & s \\ s & \# & v & o & \longrightarrow' & v \quad o \end{array}$$

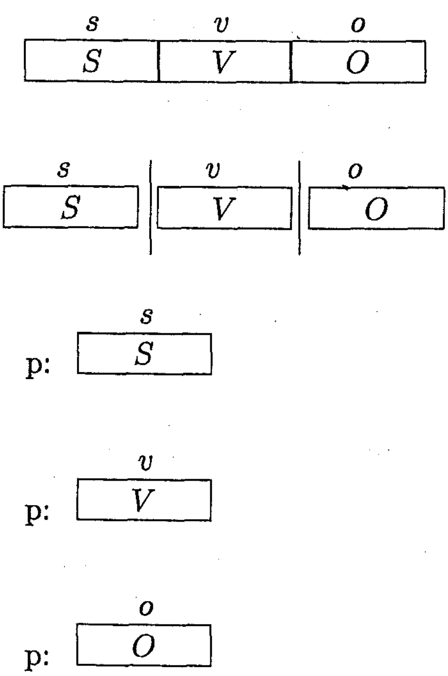
If we summarize the pieces that we have produced by means of these two cuts, we realize that we have obtained all the possible combinations with simple strings except one:

$$\begin{pmatrix} v \\ V \end{pmatrix}$$

Relying on this piece may be important for further operations, but obviously it cannot be a basic piece because we have to achieve it by means of a complex cut, that is, it cannot be produced with a single cut in a basic ULPS. Therefore, we will introduce it within the inventory of pieces that may be achieved by cutting a basic ULPS.

$$\text{Cut}\begin{pmatrix} v & \# & o \\ V & \# & O \end{pmatrix} + \text{Cut}\begin{pmatrix} s & \# & v \\ S & \# & V \end{pmatrix}$$

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008



The formal representation is:

$$\begin{pmatrix} s & \# & v & \# & o \\ S & \# & V & \# & O \end{pmatrix} \Longrightarrow \begin{pmatrix} s \\ S \end{pmatrix}, \begin{pmatrix} v \\ V \end{pmatrix}, \begin{pmatrix} o \\ O \end{pmatrix}$$

6.1.3 Staggered cuts

There are three kinds of staggered cuts:

1. Focal cuts.

2. Variable cuts.

3. Internal cuts.

As we have pointed in 6.1.1, we can only accept cuts with resulting pieces with, at least, one variable and one focus. We are not interested in those fragments without variables or focus because the operations we can bring about

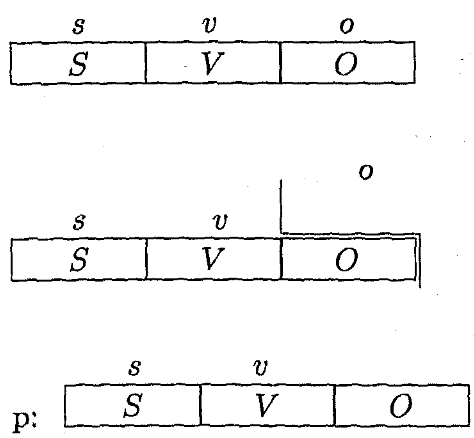
ISBN:978-84-691-1896-2/D.L:T-352-2008

with them are analogous to those we have carried out with blunt cuts. We exclude cuts $\begin{pmatrix} \# & s & v & o \\ S & V & O & \# \end{pmatrix}$ and $\begin{pmatrix} s & v & o & \# \\ \# & S & V & O \end{pmatrix}$ because of the above reason.

Cuts that correspond to (1) and (2) will produce only one piece, whereas cuts belonging to (3) will produce two pieces.

6.1.3.1 Focal cuts

a) Cut $\begin{pmatrix} \# & o \\ O & \# \end{pmatrix}$

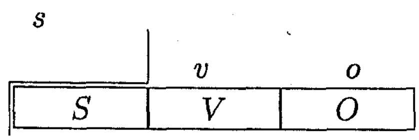


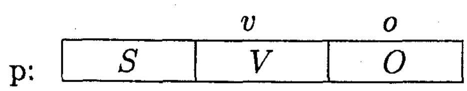
Formal representation:

$$\begin{pmatrix} s & v & \# & o \\ S & V & O & \# \end{pmatrix} \longrightarrow \begin{pmatrix} s & v & & \\ S & V & O & \end{pmatrix}$$

b) Cut $\begin{pmatrix} s & \# \\ \# & S \end{pmatrix}$

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

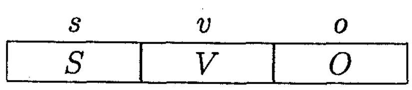


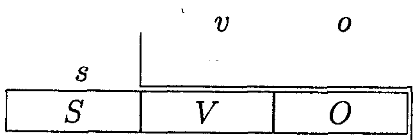


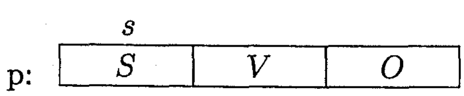
Formal representation:

$$\begin{pmatrix} s & \# & v & o \\ \# & S & V & O \end{pmatrix} \longrightarrow \begin{pmatrix} & v & o \\ S & V & O \end{pmatrix}$$

c) Cut $\begin{pmatrix} \# & v & o \\ V & O & \# \end{pmatrix}$





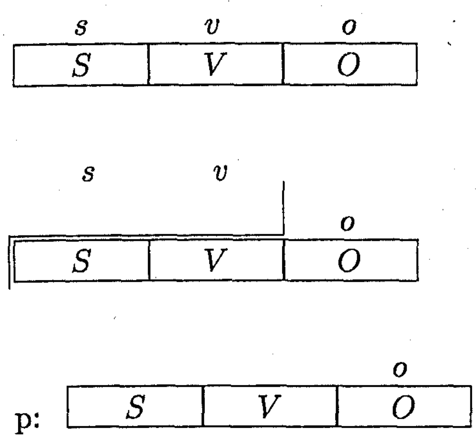


The formal representation is:

$$\begin{pmatrix} s & \# & v & o \\ S & V & O & \# \end{pmatrix} \longrightarrow \begin{pmatrix} s & & \\ S & V & O \end{pmatrix}$$

d) Cut $\begin{pmatrix} s & v & \# \\ \# & S & V \end{pmatrix}$

ISBN:978-84-691-1896-2/D.L:T-352-2008

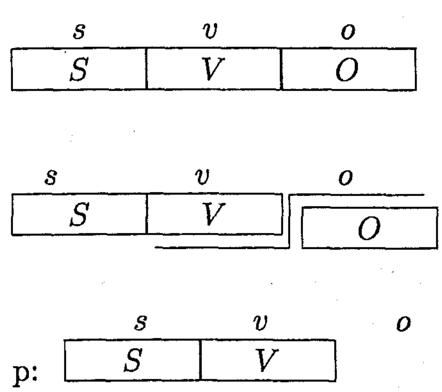


The formal representation is:

$$\begin{pmatrix} s & v & \# & o \\ \# & S & V & O \end{pmatrix} \longrightarrow \begin{pmatrix} & & o \\ S & V & O \end{pmatrix}$$

6.1.3.2 Variable cuts

a) Cut $\begin{pmatrix} o & \# \\ \# & O \end{pmatrix}$

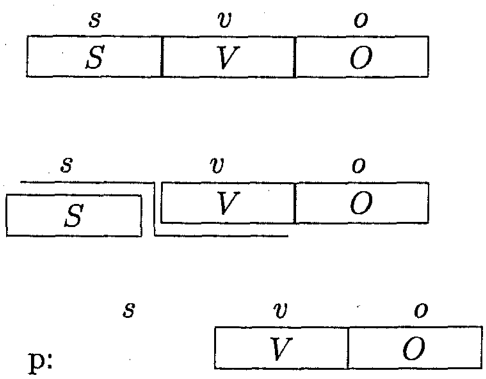


Formal representation:

$$\begin{pmatrix} s & v & o & \# \\ S & V & \# & O \end{pmatrix} \longrightarrow \begin{pmatrix} s & v & o \\ S & V & \end{pmatrix}$$

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

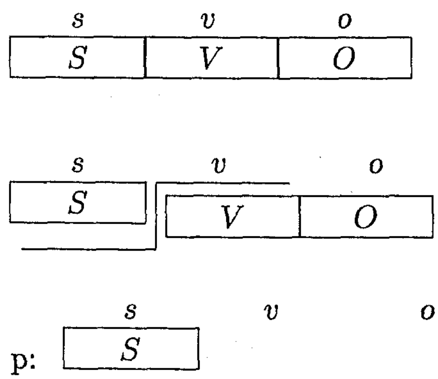
b) Cut $\begin{pmatrix} \# & s \\ S & \# \end{pmatrix}$



Formal representation:

$$\begin{pmatrix} \# & s & v & o \\ S & \# & V & O \end{pmatrix} \longrightarrow \begin{pmatrix} s & v & o \\ & V & O \end{pmatrix}$$

c) Cut $\begin{pmatrix} v & o & \# \\ \# & V & O \end{pmatrix}$

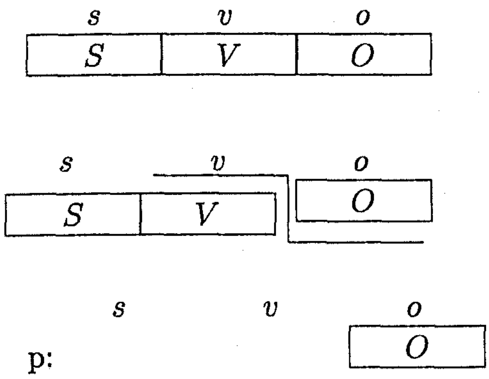


The formal representation is:

$$\begin{pmatrix} s & v & o & \# \\ S & \# & V & O \end{pmatrix} \longrightarrow \begin{pmatrix} s & v & o \\ S & & \end{pmatrix}$$

ISBN:978-84-691-1896-2/D.L:T-352-2008

d) Cut $\begin{pmatrix} \# & s & v \\ S & V & \# \end{pmatrix}$

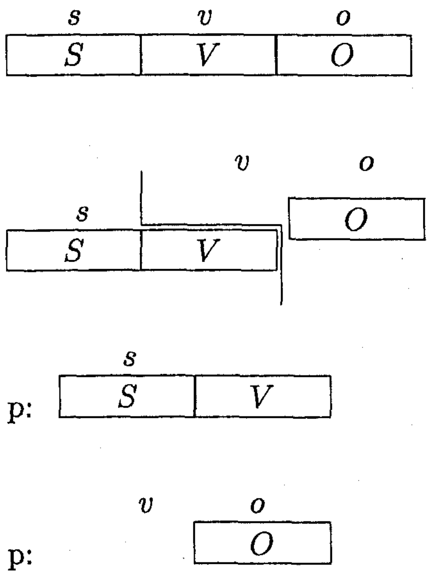


The formal representation is:

$$\begin{pmatrix} \# & s & v & o \\ S & V & \# & O \end{pmatrix} \longrightarrow \begin{pmatrix} s & v & o \\ & & O \end{pmatrix}$$

6.1.3.3 Internal cuts

a) Cut $\begin{pmatrix} \# & v \\ V & \# \end{pmatrix}$**:**

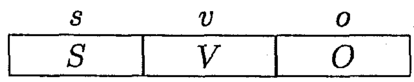


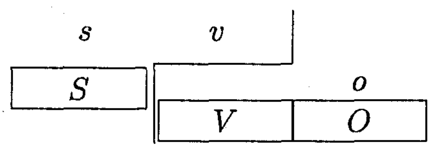
UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

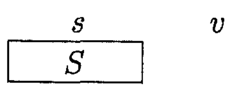
The formal representation is:

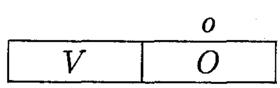
$$\begin{pmatrix} s & \# & v & o \\ S & V & \# & O \end{pmatrix} \Longrightarrow \begin{pmatrix} s & \\ S & V \end{pmatrix}, \begin{pmatrix} v & o \\ & O \end{pmatrix}$$

b) Cut $\begin{pmatrix} v & \# \\ \# & V \end{pmatrix}$:





p: 

p: 

Formal representation:

$$\begin{pmatrix} s & v & \# & o \\ S & \# & V & O \end{pmatrix} \Longrightarrow \begin{pmatrix} s & v \\ S & \end{pmatrix}, \begin{pmatrix} & o \\ V & O \end{pmatrix}$$

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

6.1.4 Syntactic pieces

We have obtained 17 pieces by means of all the possible simple, blunt or staggered cuts carried out in a basic string:

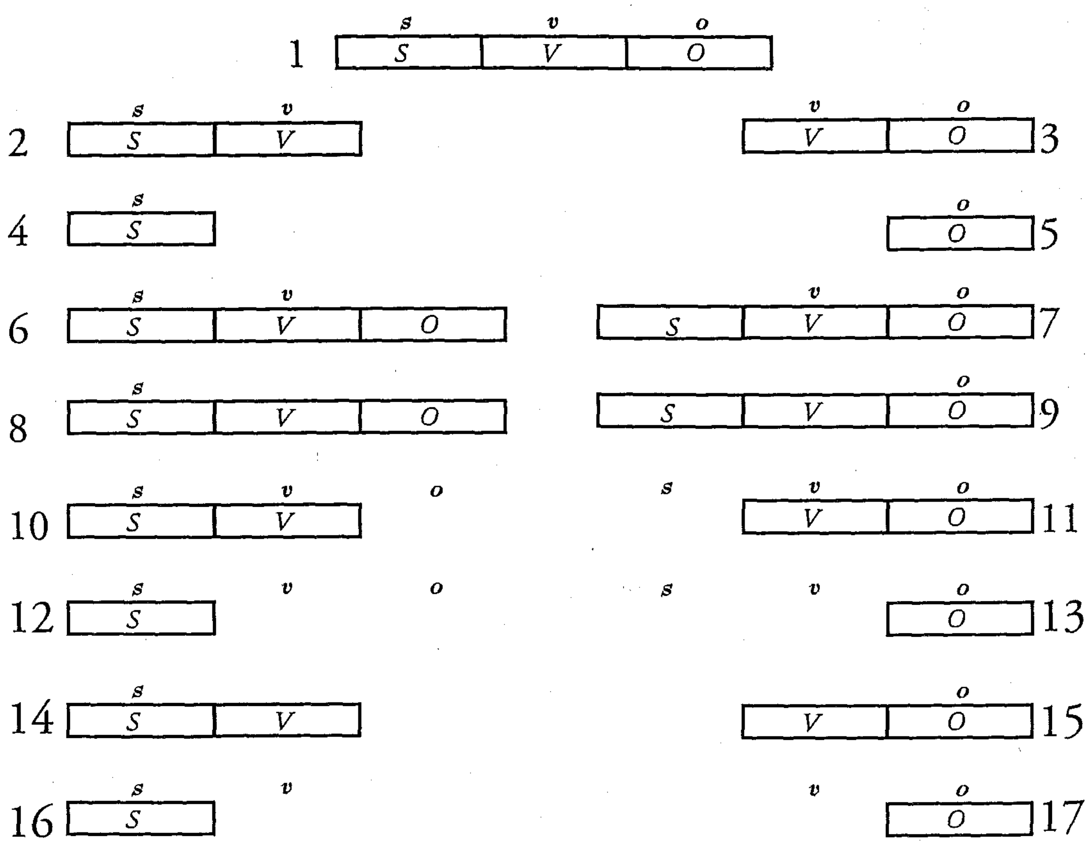


Figure 6.1: Syntactic pieces

whose formal representation is:

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

Piece 1 $\begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix}$

Piece 2 $\begin{pmatrix} s & v \\ S & V \end{pmatrix}$ Piece 3 $\begin{pmatrix} v & o \\ V & O \end{pmatrix}$

Piece 4 $\begin{pmatrix} s \\ S \end{pmatrix}$ Piece 5 $\begin{pmatrix} o \\ O \end{pmatrix}$

Piece 6 $\begin{pmatrix} s & v & \\ S & V & O \end{pmatrix}$ Piece 7 $\begin{pmatrix} & v & o \\ S & V & O \end{pmatrix}$

Piece 8 $\begin{pmatrix} s & & \\ S & V & O \end{pmatrix}$ Piece 9 $\begin{pmatrix} & & o \\ S & V & O \end{pmatrix}$

Piece 10 $\begin{pmatrix} s & v & o \\ S & V & \end{pmatrix}$ Piece 11 $\begin{pmatrix} s & v & o \\ & V & O \end{pmatrix}$

Piece 12 $\begin{pmatrix} s & v & o \\ S & & \end{pmatrix}$ Piece 13 $\begin{pmatrix} s & v & o \\ & & O \end{pmatrix}$

Piece 14 $\begin{pmatrix} s & \\ S & V \end{pmatrix}$ Piece 15 $\begin{pmatrix} & o \\ V & O \end{pmatrix}$

Piece 16 $\begin{pmatrix} s & v \\ S & \end{pmatrix}$ Piece 17 $\begin{pmatrix} v & o \\ & O \end{pmatrix}$

6.1.5 Results

In total, we have carried out 12 simple cuts in the *basic ULPS* achieving the following results:

ISBN:978-84-691-1896-2/D.L:T-352-2008

- Two blunt cuts that have produced four syntactic pieces.
- Ten staggered cuts that have produced twelve syntactic pieces.

On the whole and taking the basic ULPS into account, we have achieved 17 syntactic pieces in total, as it is indicated in figure 5.7. Five of them have blunt ends, whereas 12 have one staggered end. There is not any syntactic piece achieved by a simple cut with overhangs at both ends.

Therefore, we can conclude by saying that blunt cuts are much more yielding because a) all the pieces that they produce may be used for making new linguistic structures and b) every blunt cut produces two pieces.

6.2 Links

6.2.1 Objectives

In this section we want to check all kinds of existing links, whether they match the kinds of cuts and what the result of each one of them is. In order to do this, we will use the 17 syntactic pieces we have achieved in the previous section.

Since the cuts carried out are of two kinds: a) blunt ones, b) staggered ones, then three kinds of links may exist, at most:

1. A link of pieces with blunt ends.
2. A link of pieces with staggered ends, which can be at the same time:
   (a) A link of pieces with overhangs of 1 in depth.
   (b) A link of pieces with overhangs of 2 in depth.
3. A link of blunt ends pieces with staggered ends pieces.

The objective is to remove those pieces that did not achieve satisfactory results in their links, which also means that we can probably further limit the set of cuts.

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

6.2.2 Link of blunt ends: splicing

We can make any desirable combination with blunt ends of any piece because, in principle, the link is arbitrary. We will only require the presence of some element capable of catalysing it, as we will see later. Let's show some examples:

Example 1: Piece 2 - Piece 10

s	*v*	*s*	*v*	*o*
S	*V*	*S*	*V*	

Example 2: Piece 1 - Piece 3

s	*v*	*o*	*v*	*o*
S	*V*	*O*	*V*	*O*

Example 3: Piece 11 - Piece 8

	v	*o*	*s*		
S	*V*	*O*	*S*	*V*	*O*

Figure 6.2: Arbitrary link of pieces with blunt ends

Splicing does not have any restriction to theoretical level. We need to take into account the agreement and the level assignment criteria when combining syntactic pieces with specific focus in order to distinguish which foci can be combined and which cannot.

However, this method has a problem: it is too powerful. Due to this problem, we will not ponder all the possible cases in the section in which we will refer to splicing, but we will make a restriction by following theoretical bases established by (Head, 1987) and later by (Păun, Rozenberg & Salomaa, 1998).

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

6.2.3 Link of staggered ends: sticker link

Pieces with overhangs composed by variables or by foci can be stuck as in DNA molecules, which have been cut by an enzyme. In order to fit these molecules in the DNA, there are two conditions that must be achieved: a) overhangs must be the same depth, b) bases must be complementary. Since, as we have said in chapter 3, it seems that there is not complementarity in linguistic structures, we can try to fit pieces with overhangs of the same depth in different levels.

We start by sticking pieces with overhangs of 1 depth (figure 6.3).

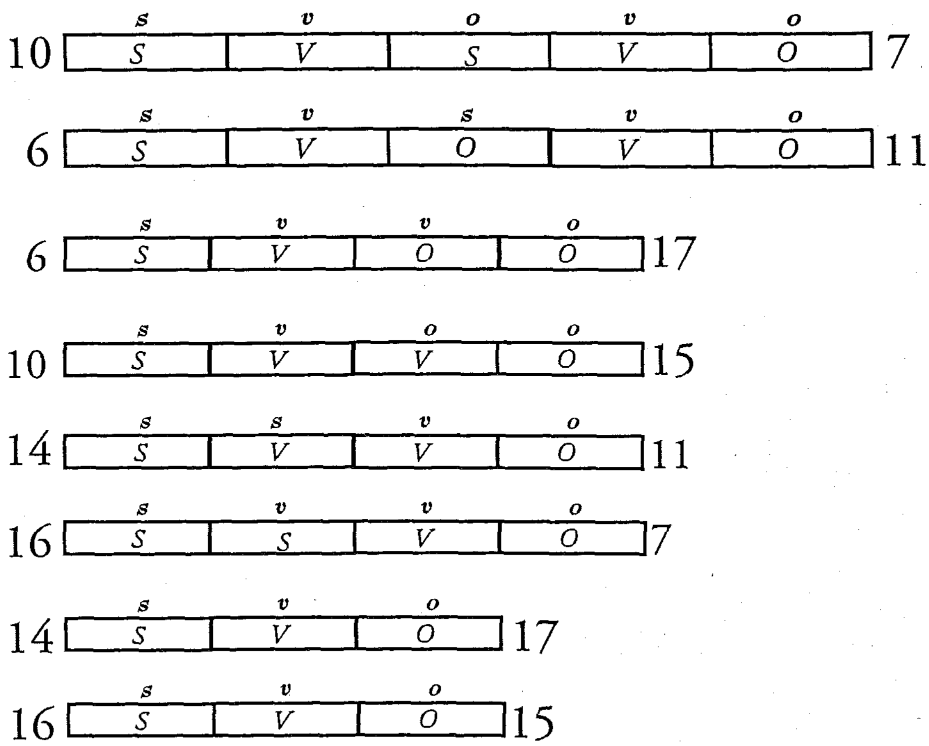


Figure 6.3: Links of pieces with staggered cuts of 1 in depth

We can carry out 8 links in total with the units we have. However, it is not

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

certain that all of them are linguistically possible. Looking at each of them, we realize that 6 from 8 have a pole with unknown characteristics until now. These poles are formed by one variable and one focus which do not correspond to its domain. These are the following ones:

- Link 10 – 7: pole $\begin{pmatrix} o \\ \mathcal{S} \end{pmatrix}$
- Link 6 – 11: pole $\begin{pmatrix} s \\ \mathcal{O} \end{pmatrix}$
- Link 6 – 17: pole $\begin{pmatrix} v \\ \mathcal{O} \end{pmatrix}$
- Link 10 –15: pole $\begin{pmatrix} o \\ \mathcal{V} \end{pmatrix}$
- Link 14 – 11: pole $\begin{pmatrix} s \\ \mathcal{V} \end{pmatrix}$
- Link 16 – 7: pole$\begin{pmatrix} v \\ \mathcal{S} \end{pmatrix}$

The appearance of these poles seems to contradict the pattern definition in itself, and it supposes a serious transgression of the fundamental principle of formation of ULPS. Consequently, it is necessary to establish the authenticity of these junctions.

We can find the answer to the proposed problem in the description of the domains of each variable. We have assured that $\mathcal{S}$ and $\mathcal{V}$ are disjoint, as well as $\mathcal{V}$ and $\mathcal{O}$. Nevertheless, $\mathcal{S}$ and $\mathcal{O}$ are not disjoint, which means that there are elements belonging to $\mathcal{S}$ that also belong to $\mathcal{O}$. We call $\mathcal{I}$ to $\mathcal{S} \cap \mathcal{O}$. It follows that there can be apparently bad-shaped poles which are acceptable in

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

their syntactic structure; concretely, those in which variables S and O play a part, those which share a part of the domain.

What we have just exposed suggests that there is also some kind of complementarity in the conformation of ULPS which we call *polar coherence.*

Let's establish how this criterion of coherence works, that is to say, with the focus of which domain a specific variable can be replaced. When a pattern and a focus can be joined, we denote it by means of a$\rightleftharpoons$ sign, and when they cannot be joined we denote it by means of a $\asymp$. According to the domain's definition, it is steady that: $S \rightleftharpoons s$, $S \asymp v$, $V \asymp s$, $V \asymp o$, $V \rightleftharpoons v$, $O \rightleftharpoons o$, $O \asymp v$, $S \rightleftharpoons o$ iff $o \in \mathcal{I}$, and $O \rightleftharpoons s$ iff $s \in \mathcal{I}$.

Thus, the polar coherence rule for every variable establishes that:

- $S \rightleftharpoons z|z \in \mathcal{S}, z \in \mathcal{I}$; $S \asymp z|z \in \mathcal{V}$
- $O \rightleftharpoons z|z \in \mathcal{O}, z \in \mathcal{I}$; $S \asymp z|z \in \mathcal{V}$
- $V \rightleftharpoons z|z \in \mathcal{V}$; $V \asymp z|z \in \mathcal{S}, z \in \mathcal{O}$

This is a theoretical approach that refers to 'possibility', because the adjustment of a string belonging to $\mathcal{S}$ or to $\mathcal{O}$ within a ULPS depends always on the conditions imposed by v concerning the agreement and level assignment. So, $S \rightleftharpoons o$ means that "*it is possible*" that this variable and a focus of the domain of $\mathcal{O}$ may be coherent, whereas $S \asymp v$ means that "*it is not possible*" that variable S and a focus of the domain of $\mathcal{V}$ are coherent. To sum up, the negative form means "never", the positive form means "sometimes".

From what we have said up to now now, it follows that there are two kinds of poles:

- **Coherent** poles: they are **possibles** in well-shaped ULPS. They are made up of one variable and one string that are in agreement with the principle of coherence: $\mathcal{Z} \rightleftharpoons z$. They are the poles: $\begin{pmatrix} s \\ S \end{pmatrix}$, $\begin{pmatrix} v \\ V \end{pmatrix}$,

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

$$\begin{pmatrix} o \\ O \end{pmatrix}, \begin{pmatrix} s \\ O \end{pmatrix}, \begin{pmatrix} o \\ S \end{pmatrix}.$$

- **Incoherent** poles: they are **impossible** in a well-shaped ULPS. They are made up of one variable and one string that are not in agreement with the principle of coherence: $\mathcal{Z} \asymp z$. These are the $\begin{pmatrix} s \\ V \end{pmatrix}$, $\begin{pmatrix} v \\ S \end{pmatrix}$, $\begin{pmatrix} o \\ V \end{pmatrix}$, $\begin{pmatrix} v \\ O \end{pmatrix}$ poles.

The *polar coherence* criterion makes us think, once more, about the analogy between the recombination methods used by the genetic language and those used by the verbal one.

So, in order to join sticker ends, it has not been necessary to modify any of the two conditions imposed by the DNA fragments to be joined. It has only been necessary to adjust the complementarity theory to linguistic pieces. If we want to join two linguistic fragments with staggered ends it is necessary that:

- they have overhangs of the same depth in the opposite direction,
- variables and focus of the overhangs are coherent.

According to these criteria, links 6 – 7, 10 – 15, 15 – 11, 16 – 7 must be excluded. Consequently, from the eight pieces made by means of joining staggered ends, four must be discarded, leaving only 10 – 7, 6 – 11, 14 – 17 and 16 –15. According to the 14 – 17 and 16 –15 links, they are not acceptable but also *tautological*. We call them in this way because the pieces they form can be achieved by means of only one cut in a basic ULPS.

We still must test pieces with overhangs in depth 2. There are only two possible links with them.

Link 8 – 13 has the $\begin{pmatrix} s \\ V \end{pmatrix}$ and $\begin{pmatrix} v \\ O \end{pmatrix}$ poles, which are not coherent. Link

ISBN:978-84-691-1896-2/D.L:T-352-2008

Piece 8 - Piece 13

s	*s*	*v*	*o*
S	*V*	*O*	*O*

Piece 12 - Piece 9

s	*v*	*o*	*o*
S	*S*	*V*	*O*

Figure 6.4: Links of pieces with staggered cuts of 2 in depth

12 − 9 holds $\begin{pmatrix} v \\ S \end{pmatrix}$ and $\begin{pmatrix} o \\ V \end{pmatrix}$ poles, which are also incoherent. Therefore, we can conclude by saying that the link of two pieces with overhangs in depth 2 is not possible if we deal with fragments that have been achieved by means of the cut of basic ULPS. We assume the methodological limitation of working with basic initial patterns, which puts us under the constraint of rejecting the possibility of working with cuts of 2 in depth.

6.2.4 Link of level

Finally, we consider a different way of joining these strings, which is unknown in the genetic process and that could entail, therefore, some kind of linguistic innovation. Once here, it is worthwhile to propose another kind of link: the *link of level*. It is produced between the staggered end of a ULPS and another one with both blunt ends.

In order to achieve a group which does not increase in a geometrical way, we impose some conditions to this kind of cuts that later will be justified by means of the proofs that will be given in this section:

(i) The ULPS with staggered end must have been submitted to one cut of

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

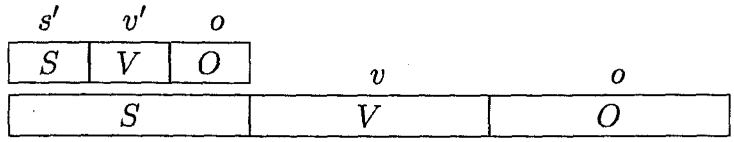
focus. That is to say, only pieces with a variable in the overhang are acceptable.

(ii) The ULPS with blunt ends must contain focus v.

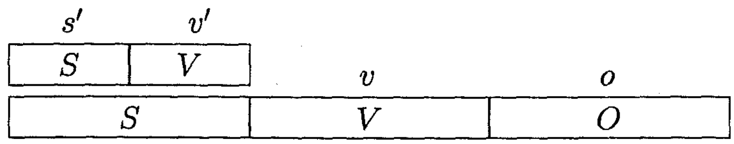
(iii) The ULPS with staggered end must have the overhang in depth 1.

According to the pieces that we have and taking into account conditions (i), (ii), (iii) the possible links of level are the following:

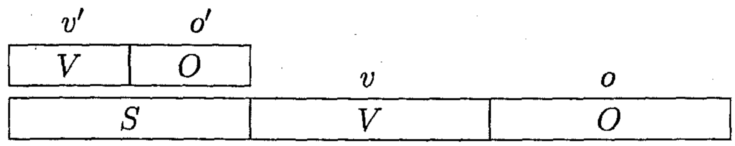
a. Piece 1 – Piece 7



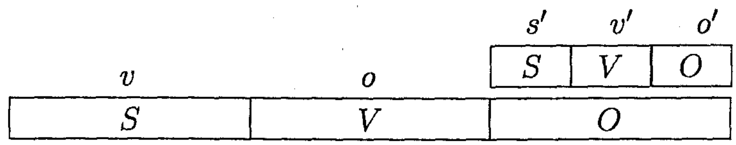
b. Piece 2 – Piece 7



c. Piece 3 – Piece 7

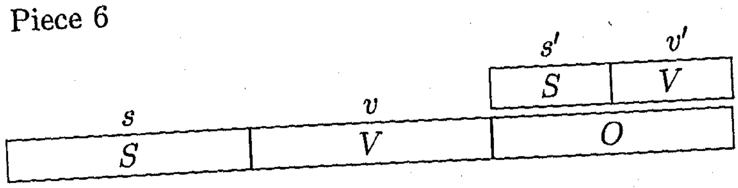


d. Piece 1 – Piece 6



UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

e. Piece 2 – Piece 6



f. Piece 3 – Piece 6

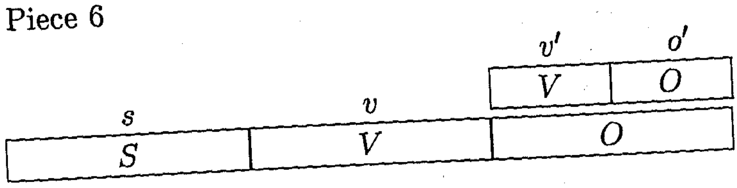


Figure 6.5: Link of level

It might seem that the links we have just performed do not follow the principle of *polar coherence*, but it is not the case. In the above examples a new kind of pole has been generated. It does not have two levels, as we have seen until now, but three. The first pole of a) must be represented in this way $\begin{pmatrix} s & v & o \\ S & V & O \\ & S & \end{pmatrix}$. The polar coherence must be established among the elements of the first two levels - which fulfill it - and not of the third one. Every level can be given a role: $\begin{pmatrix} focus \\ functionalvar. \\ function \end{pmatrix}$. Functions follow from the variables. Patterns do not have function. However, the last level gives functionality to the attached pattern. When we will study this kind of links, we will see specific examples.

Let's present now a formal representation of the 6 previous links:

a. Piece 1 – Piece 7

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

a) $s_{PN}v_{PN}o$ $s'v'o'$
b) $(\ )_{PN}v_{PN}o$ $s'v'o'$
c) $(s'v'o')_{PN}v_{PN}o$

b. Piece 2 – Piece 7

a) $s_{PN}v_{PN}o$ $s'v'$
b) $(\ )_{PN}v_{PN}o$ $s'v'$
c) $(s'v')_{PN}v_{PN}o$

c. Piece 3 – Piece 7

a) $s_{PN}v_{PN}o$ $v'o'$
b) $(\ )_{PN}v_{PN}o$ $v'o'$
c) $(v'o')_{PN}v_{PN}o$

d. Piece 1 – Piece 6

a) svo_l $s'v'o'$
b) $sv(\ )_l$ $s'v'o'$
c) $sv(s'v'o')_l$

e. Piece 2 – Piece 6

a) svo_l $s'v'$
b) $sv(\ )_l$ $s'v'$
c) $sv(s'v')_l$

f. Piece 3 – Piece 6

$a)$ svo_l $\quad v'o'$
$b)$ $sv(\quad)_l$ $\quad v'o'$
$c)$ $sv(v'o')_l$

Now we can justify the stated restrictions in the link of level we saw at the beginning of this section:

(i) Impossibility of carrying out links of level with overhang pieces in the string. We have just seen that, once the link has been carried out, the polar configuration generates a

$$\begin{pmatrix} focus \\ functional\ var. \\ function \end{pmatrix}$$

structure in which the variable without focus gives function to the attached pattern. Inversely, there is not any reason to generate

$$\begin{pmatrix} focus \\ focus \\ variable \end{pmatrix}$$

structures in which a physical superposition of strings that seems impossible is shown. The function-variables, theoretical elements, can be superpositioned; strings, physical elements cannot.

(ii) Impossibility of carrying out a link of level with a string lacking in v. Variables V are, surely, the most powerful ones. The adjudication of functionality to a pattern, which we have seen before, goes by the necessary existence of V. The function of any pattern which contains V can be given or modified. The function of any pattern without V cannot. On the other hand, it is valid to ask oneself if the functionality of only one variable can be changed. For instance, whether a three level pole

ISBN:978-84-691-1896-2/D.L:T-352-2008

$$\begin{pmatrix} focus \\ variable \\ funcion \end{pmatrix}$$

could exist instead of a pole

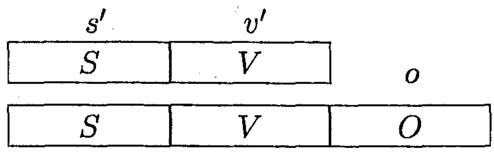
$$\begin{pmatrix} focus \\ pattern \\ function \end{pmatrix}$$

The answer is yes, as long as the variable is V. There is not any other variable capable of changing the initial functionality. Therefore, the only poles of three levels which are possible with these characteristics are:

$$\begin{pmatrix} v \\ V \\ S \end{pmatrix}, \begin{pmatrix} v \\ V \\ O \end{pmatrix}$$

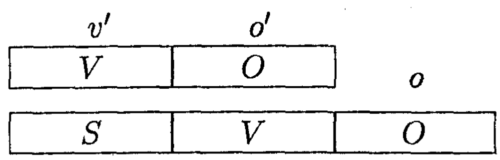
(iii) Impossibility of carrying out links of level using pieces with overhangs in depth 2. In order to join a piece in depth 2, like (9), with a piece with two poles generated by means of a blunt cut, there are two options:

- Link 9 - 12. This operation has no sense because it superimposes variable S to variable S and variable $V$to variable V, without adding anything interesting to the basic resultant string. This same effect is achieved by means of link 2-5 or 4-3 carried out with blunt cut pieces, which are much more economical. Let's remember that for equal results a staggered cut it is preferable because it has less complexity.



UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

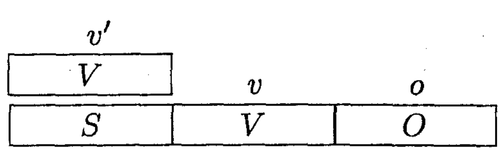
- Link 9 - 3. By means of this operation, two impossible structures are obtained due to the fact that although variables $\begin{pmatrix} V \\ S \end{pmatrix}$ are capable of being superimposed, $\begin{pmatrix} O \\ V \end{pmatrix}$ are not.



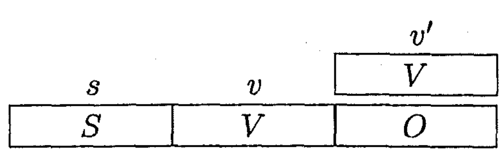
Following these three impossibilities, nothing hinders us from trying a complex operation with these pieces: a) $\begin{pmatrix} v \\ V \end{pmatrix}$ - Piece 7, b) $\begin{pmatrix} v \\ V \end{pmatrix}$ - Piece 6.

Such operations obey the three link of level rules and, therefore, we think it is good to insert them due to their efficiency when explaining some syntactic phenomena.

a) $\begin{pmatrix} O \\ V \end{pmatrix}$ - Piece 7



b) $\begin{pmatrix} O \\ V \end{pmatrix}$ - Piece 6



6.2.5 Results

When checking the unsuitability of links of pieces with an overhang of 2 in depth, the 17 initial pieces, by disposal of 8, 9, 12, 13, have been reduced

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

to 13: 1 belongs to the basic string without effecting any cut, the 4 of blunt cut, and the 8 of staggered cut of 1 in depth. Of these 13, there are 4, which have been obtained with the inside cut that can only be combined among themselves.

These 13 will be the syntactic pieces that will act as base for the operations that will take place in this thesis. We call them *basic syntactic pieces.*

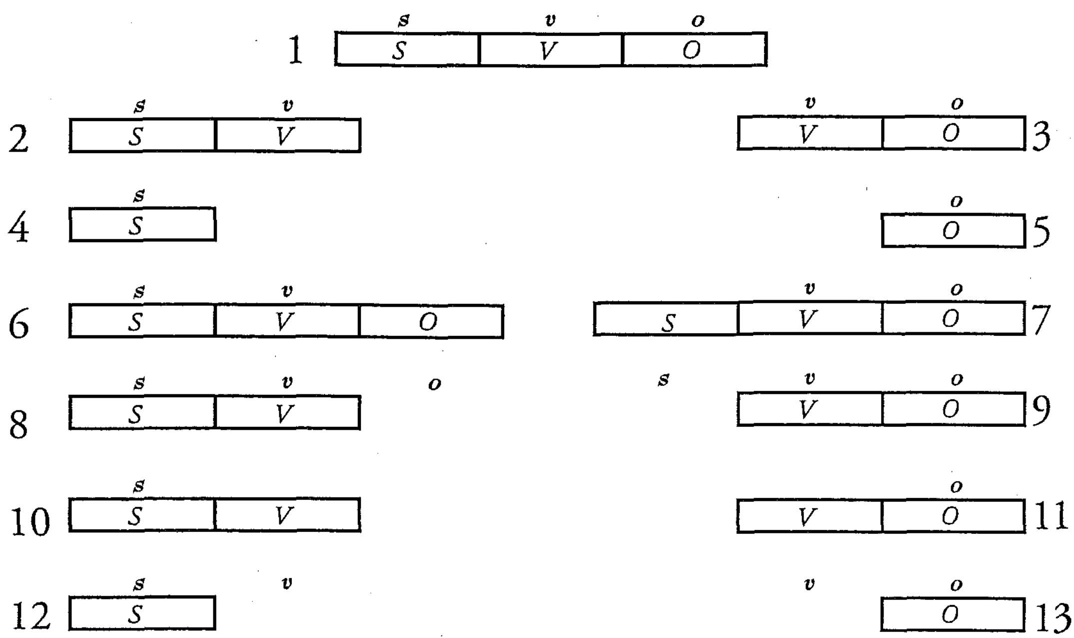


Figure 6.6: Basic syntactic pieces

We can also discard cuts 6.1.3.1 c), 6.1.3.1 d), 6.1.3.2 c) i 6.1.3.2 d) that have as a result those pieces where not one correct link can be made. Therefore, the following ones will be considered *basic cuts*:

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

T.1 Cut $\begin{pmatrix} v & \# & o \\ V & \# & O \end{pmatrix}$	T.2 Cut $\begin{pmatrix} s & \# & v \\ S & \# & V \end{pmatrix}$
T.3 Cut $\begin{pmatrix} \# & o \\ O & \# \end{pmatrix}$	T.4 Cut $\begin{pmatrix} s & \# \\ \# & S \end{pmatrix}$
T.5 Cut $\begin{pmatrix} o & \# \\ \# & O \end{pmatrix}$	T.6 Cut $\begin{pmatrix} \# & s \\ S & \# \end{pmatrix}$
T.7 Cut $\begin{pmatrix} v & \# \\ \# & V \end{pmatrix}$	T.8 Cut $\begin{pmatrix} \# & v \\ V & \# \end{pmatrix}$

Table 6.1: Basic simple cuts

On the other hand, in this section we have achieved:

- To establish a polar coherence rule which allows the exchange of variables and focus.
- To introduce a new link - the link of level - which cannot be compared to any kind of genetical link, as far as we know.
- To describe new kinds of poles : poles of three levels

Finally, we will establish a relation among the cuts and links we have tried out in this section, and the systems we will study starting from the next chapter. We use the blunt cut for:

- splicing,
- mutations: transition and transversion

We use the staggered cut for:

- sticker links,

- mixed links
- mutations: transposition and duplication.

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

Chapter 7

Ghosts

7.1 Definition and objectives

One of the essential characteristics of operations with linguistic pieces is that some elements appear in the closing strings of a derivation which were not present at the stage strings and without carrying out any kind of insertion rule. These elements are named ghosts.
We define ghosts, which we denote by f, as elements that:

(i) only appear in derivative linguistic strings,

(ii) are not present at basic strings,

(iii) have not been directly introduced by means of insertion,

(iiii) catalyse the recombination process.

We uphold that ghosts appear in the last stratum of a sentence. It is not possible to find them among variables, but only on the terminal string level.

Although the term ghost may suggest that these elements arise in a spontaneous way and by natural generation, the truth is that they have some strict rules of appearance and functioning. Therefore, we will try to characterize

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

ghosts in order to be ready when they appear at the moment of cutting and joining sentences.

First of all, we consider the existence of four kinds:

- emergent ghosts, $f \smile$,
- level assigner ghosts, $f\wr$,
- relativizator ghosts, $f \sim$,
- replicative ghosts, $f\curlyvee$.

From each one of these kinds, we will define and impose some appearance conditions depending on:

- number of necessary strings,
- kind of cut,
- phenomenon that provokes them.

We start from the premise that in every linguistic structure to which no operation has been applied the primary ULPS is also basic. That is to say, a primary string never holds ghosts.

7.2 Emergent ghosts, $f \smile$

Emergent ghosts are produced by *friction*, which takes place in the process of linking of two strings with blunt ends. We name friction the concatenation of foci belonging to the same domain or sequences of equal foci (named friction groups) in a string.

UNIVERSITAT ROVIRA I VIRGILI
MOLECULAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

7.2.1 Number of necessary strings

Emergent ghosts $f \smile$ can only arise as an outcome of operations made up with, at least, two stage strings. The reason for this is that they relate repeated foci, which never take place in a basic string where, by definition: $|s| = |v| = |o| = 1$.

7.2.2 Kind of cut

Emergent ghosts arise by means of the link of:

- pieces obtained with a blunt cut,
- basic ULPS .

Although we have limited the experimentation field to a basic ULPS in a methodological way. Any syntactic piece with a blunt end can easily be effected by the linking that produces an emergent ghost.

7.2.3 Phenomenon that causes them

Emergent ghosts are caused by *friction.*

Context or friction groups are eight if we follow the possible combinations with the application of *only one cut and link rule* with blunt or staggered cut of one in depth in basic stage strings:

1. $\begin{pmatrix} s \\ S \end{pmatrix} \begin{pmatrix} s' \\ S \end{pmatrix} \rightarrow \begin{pmatrix} s & \smile & s' \\ S & & S \end{pmatrix}$

2. $\begin{pmatrix} v \\ V \end{pmatrix} \begin{pmatrix} v' \\ V \end{pmatrix} \rightarrow \begin{pmatrix} v & \smile & v' \\ V & & V \end{pmatrix}$

3. $\begin{pmatrix} o \\ O \end{pmatrix} \begin{pmatrix} o' \\ O \end{pmatrix} \rightarrow \begin{pmatrix} o & \smile & o' \\ O & & O \end{pmatrix}$

ISBN:978-84-691-1896-2/D.L:T-352-2008

4. $\begin{pmatrix} s & v \\ S & V \end{pmatrix} \begin{pmatrix} s' & v' \\ S & V \end{pmatrix} \rightarrow \begin{pmatrix} s & v & \smile & s' & v' \\ S & V & & S & V \end{pmatrix}$

5. $\begin{pmatrix} v & o \\ V & O \end{pmatrix} \begin{pmatrix} v' & o' \\ V & O \end{pmatrix} \rightarrow \begin{pmatrix} v & o & \smile & v' & o' \\ V & O & & V & O \end{pmatrix}$

6. $\begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix} \begin{pmatrix} s' & v' & o' \\ S & V & O \end{pmatrix} \rightarrow \begin{pmatrix} s & v & o & \smile & s' & v' & o' \\ S & V & O & & S & V & O \end{pmatrix}$

According to the definition of *friction group*, these are the only possible ones. But we want now to propose the idea that the focus –and, for extension, the pole– o is a *filter focus*. o allows ghosts and ghosts' effect. Then, $f\cdot$ can join $v \cdot v$ in the same way that $vo \cdot v$. o also works as filter focus in $s \cdot s - so \cdot s$, but this is not a possible context with only the basic rules. Up to now, focus o allows ghosts to work in:

7. $\begin{pmatrix} v & o \\ V & O \end{pmatrix} \begin{pmatrix} v' \\ V \end{pmatrix} \rightarrow \begin{pmatrix} v & o & \smile & v' \\ V & O & & V \end{pmatrix}$

8. $\begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix} \begin{pmatrix} sv' \\ SV \end{pmatrix} \rightarrow \begin{pmatrix} s & v & o & \smile & s & v' \\ S & V & O & & S & V \end{pmatrix}$

Thus, we add these two contexts, which can be useful for explaining some fenomena that we will find below.

Within emergent ghosts, we can distinguish two different kinds of them according to the friction groups that cause them:

- connector ghosts $f\cdot$,
- bounder ghosts $f\star$.

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

7.2.4 Connector ghosts $f\cdot$

Connectors are universal emergent ghosts. They simply link two equal things. They are assimilated to an addition rule because they only pile elements. As a general rule, we may say that the appearance of a connector ghost between two friction groups is always possible in blunt cut systems. This is to say, the appearance of $f\cdot$ is possible in the interaction of the eight possible friction contexts for emergent ghosts. In spite of this universality, connector ghosts bear very strict readjustment or string selection rules in the strings, which we will see when we study the systems that cause them and rather bind their freedom of appearance.

7.2.5 Bounder ghosts $f\star$

The bounders are restricted emergent ghosts that act according to one condition: the existence of sv in the friction group. From the eight contexts where it is possible to find $f \smile$, the condition that we have just stated leaves out (1), (2), (3), (5). Therefore, the valid ones are:

4. $\begin{pmatrix} s & v \\ S & V \end{pmatrix}\begin{pmatrix} s' & v' \\ S & V \end{pmatrix} \rightarrow \begin{pmatrix} s & v & \smile & s' & v' \\ S & V & & S & V \end{pmatrix}$

6. $\begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix}\begin{pmatrix} s' & v' & o' \\ S & V & O \end{pmatrix} \rightarrow \begin{pmatrix} s & v & o & \smile & s' & v' & o' \\ S & V & O & & S & V & O \end{pmatrix}$

In addition to this, we establish another condition in order that $f\star$ really appears in (4):

1. $|\ell(v)| > 1$

2. $\exists \varnothing \in \ell(v)$

We will study more deeply the functioning and linguistic meaning of emergent ghosts in chapter 8.

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

7.3 Level assigner ghosts $f\wr$

Level assigner ghosts $f\wr$ are those that arise by friction through the application of a linking level rule.

7.3.1 Number of necessary strings

It is two because the objective of $f\wr$ is inserting a ULPS inside another one.

7.3.2 Kind of cut

One of the linking strings must have a staggered cut of 1 in depth, whereas the other one must have blunt ends or must be an axiomatic shortened ULPS. That is the reason why systems in which they arise are called mixed systems.

7.3.3 Phenomenon that causes them

Level assignment ghosts are caused by a *focal insertion*, this is, the insertion of a whole ULPS in the focal site of a pole shortned by a staggered cut.

These ghosts have only one prerequisit for arising: the inserted ULPS must have the pole $\begin{pmatrix} s \\ S \end{pmatrix}$

The functioning and linguistic meaning of level assigner ghosts will be carefully studied in chapter 11

7.4 Relativizator ghosts $f\sim$

These are caused by the link of two ULPS in which a staggered cut has been applied by the recognition of an equal focus.

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

7.4.1 Necessary strings

Two strings with some equal focus that could be $s = s'$, $s = o'$, $o = s'$, $o = o'$.

7.4.2 Kind of cut

This type is staggered by recognition. The cut must be 1 in depth. Such a cut works in the following way: there are two ULPS X and Υ. If any focus of X is equal to any focus of Υ then this focus of Υ disappears. The difference between these cuts and those which we have seen until now is that the previous ones were arbitrary and these are not.

7.4.3 Phenomenon that causes them

Disappearance of an element by recognition of an equal focus. According to the imposed conditions, the following cuts are possible:

$$1.\ \begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix}\begin{pmatrix} s' & v' & o' \\ S & V & O \end{pmatrix} \rightarrow \begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix} \sim \begin{pmatrix} & v' & o' \\ S & V & O \end{pmatrix}$$

$$2.\ \begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix}\begin{pmatrix} s' & v' & o' \\ S & V & O \end{pmatrix} \rightarrow \begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix} \sim \begin{pmatrix} s' & v' & \\ S & V & O \end{pmatrix}$$

1 can be for $s = s'$ or for $o = s'$ and 2 can be for $s = o'$ and for $o = o'$.

$f \sim$ causes the link of recognized poles by means of cutting a ULPS in order to stick them exactly where recognition has been caused.

UNIVERSITAT ROVIRA I VIRGILI
MOLECUAR COMPUTING METHODS FOR NATURAL LANGUAGE SYNTAX
Gemma Bel Enguix
ISBN:978-84-691-1896-2/D.L:T-352-2008

Example 4

For $s = s'$

$$\begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix}\begin{pmatrix} s' & v' & o' \\ S & V & O \end{pmatrix} \rightarrow \begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix} \sim \begin{pmatrix} & v' & o' \\ S & V & O \end{pmatrix} \rightarrow$$

$$\begin{pmatrix} s & \sim & v' & o' & v & o \\ S & S & V & O & V & O \end{pmatrix}$$

Functioning and linguistic meaning of relativizator ghosts will be carefully studied in chapter 10.

7.5 Replicative ghosts $f\Upsilon$

Replicative ghosts are caused by the disappearance of an element from its original place within a string. It can happen by means of deletion or by transposition through a staggered cut.

7.5.1 Number of necessary strings

One. When the ghost arises due to a mutation, it is enough with one stage string. It is only necessary that a focus is deleted, it is copied or its order is changed within its string. This is the reason why it is not necessary to formulate any interaction rule with any other structure.

7.5.2 Kind of cut

Staggered cut. When these same operations which we have just named occur with a blunt cut, ghosts do not arise. In such case, the ghost indicates that there is no focus but variable still remains.

We agree to take into account that the staggered cut cannot be an incision. This is to say, it cannot be caused by means of a cut with this structure:

ISBN:978-84-691-1896-2/D.L:T-352-2008

$$\begin{pmatrix} s & & o \\ S & V & O \end{pmatrix}$$

It invalidates focus belonging to $\mathcal{V}$ as mutant elements.

7.5.3 Phenomenon that causes them

Replicative ghosts $f\Upsilon$ are caused by any of these two kinds of mutations within a string: a) deletion, b) transposition. However, not all the mutant foci leave a ghost as a memory.

Foci v have already been deleted as mutant and, therefore, as replicative (mutant foci with ghost). Foci s seldom causes the appearance of replicative elements when they disappear. So, we establish that only foci $z|z \in \mathcal{O}$, are strongly replicative.

Example 5 *Deletion of s:*

$$\begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix} \rightarrow \begin{pmatrix} & v & o \\ S & V & O \end{pmatrix}$$

Example 6 *Deletion of o:*

$$\begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix} \rightarrow \begin{pmatrix} s & v & \\ S & V & O \end{pmatrix} \rightarrow \begin{pmatrix} s & v & \Upsilon \\ S & V & O \end{pmatrix}$$

Example 7 *Movement of o:*

$$\begin{pmatrix} s & v & o \\ S & V & O \end{pmatrix} \rightarrow \begin{pmatrix} s & v & \\ S & V & O \end{pmatrix} \rightarrow \begin{pmatrix} os & v & \Upsilon \\ S & V & O \end{pmatrix}$$