UNITS 2, 3 AND 4: LEXICAL ANALYSIS AND GRAMMAR DESIGN FOR THE SYNTAX ANALYSIS

We want to develop an analyzer that can verify that the data packages that circulate through an information channel have the appropriate structure.

A data package consists of two types of blocks, the blocks labeled with the letter "a" and those labeled with the letter "b". Valid data packages have the following structure:

- They start with a block "a"
- They can have any number of blocks greater than or equal to 1.
- They cannot have more than two consecutive blocks with the same label.

Example:

```
a abbabaab aabaa abaaab ababaabb baab ...
```

Error package (three consecutive "a" blocks)

Error package (it starts with a "b" block)

1. Define the grammar G of the analyzer.

David Griol Barres, Antonio Berlanga de Jesús, Jesús García Herrero, Juan Manuel Alonso Weber

Solution:

The problem of generating the grammar that will be used to perform the parsing can be approached from two points of view:

- Make a grammar that accepts only the valid words of the language. It has the advantage of not requiring any additional control, but the drawback of making recovery of the error more difficult and giving information about it.
- Construct a more general grammar that accepts valid and erroneous sentences, later by means of actions in the production rules, the syntactic analysis must be performed. It has the advantage of facilitating the recovery of errors and allows giving more information about the reason for the error.

Both approaches are valid, in this solution the second one will be developed.

A general grammar, the symbol E is the axiom, which accepts valid information packets (also accepts erroneous) is:

E::= S d E | S
S::=a S | b S | a | b

Factoring:

E::= S E'
E'::= d, E | λ
S::= a R | b T
R::= S | λ
T::= S | λ

The production rules for R and T are the same and can be simplified:

E::= S E'
E'::= d E | λ
S::= a R | b R
R::= S | λ

The inclusion of the token "d," allows to separate the packages, it is the token that represents the blank spaces.