

# On the difference between formal and natural languages

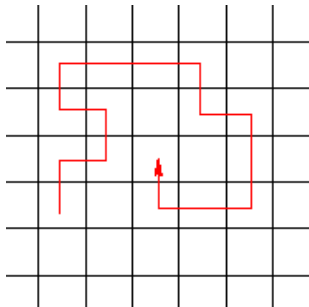
Gilles Dowek

## I. Languages in informatics





## A typical exercise in informatics



*FRFLFLFRFRFFFRFLFRFFRFFRF*

## Another typical exercise

Write a program to **translate** texts from one language to the other

# Unlike any other science

A notion of **language**

Learn a language

- ▶ Invent a language
- ▶ Translate a text from one language to another
- ▶ Improve a language  $EEE \rightarrow 3E$

# Why not using a natural language to describe this motion?

Because texts used by humans **and machines**

Machines do not understand natural languages

History of programming languages: closer and closer to natural languages: machine language, Fortran, etc. preserving possibility to be used by a machine

**History leads to relativize this idea**



II. What is a formal language?

# Specialization

Express: a command given to a robot, an algorithm, a contract, a novel, etc.

A natural language can express a command given to a robot, an algorithm, a contract, a novel, etc. (sometimes with lack of preciseness)

Natural languages are **universal**, formal languages are **specialized**

# Vocabulary

Very **limited**: *N, S, E, W*

A natural language: several thousands

But **extensible**:

*let square = NNNEESSWWW; EEE square*  
scoping, etc.

Not in natural languages:

*let poodle = feijoada; I want to eat a poodle*

# Grammar

## Simple

$$I = N \mid S \mid E \mid W$$
$$L = \varepsilon \mid I L$$

A programming language: a few pages

Portuguese / French: several hundred pages, irregular verbs, etc.

But **more precise**: a comma forgotten incorrect program

A comma forgotten in a Portuguese sentence, more or less ok

**Strict** border between correctness and incorrectness

What's up? It's awesome that you came to visit!

not as correct as

Good morning. We appreciate your visit.

but still ok

III. What are Chomsky grammars useful for?

# Correctness

requires a precise definition of the grammars

A formal language to express grammars

Backus-Naur form, Chomsky's grammar

$$I = N \mid S \mid E \mid W$$

$$L = \varepsilon \mid I L$$

## Chomsky's (initial) project

A formal language to describe the grammar of **natural** languages  
Linguists have mixed views about this project

An formal language to describe the grammar of **formal** languages  
An **excellent** tool

Serendipity

#### IV. An history of the notion of language



## A bedtime story?

Emergence of formal languages: side effect of invention of computer

Need to communicate with machines in a **common** language

**Once a concept is identified**

seek its first occurrences in history

Before the computer? ancient Greece? Phoenicia? Mesopotamia?

## Around 3300 BC

There are 237 inhabitants in this village

Discontinuity

237 an expression of another language inserted in the English text

Even a different alphabet

But not another natural language: a formal language

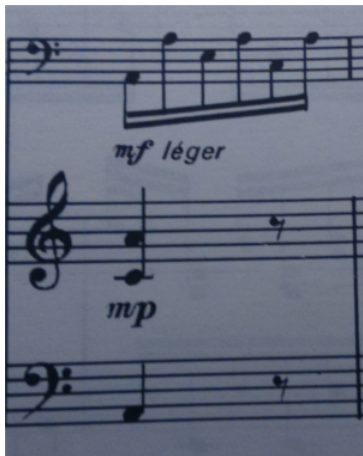
Specialized, limited vocabulary, simple grammar

$$D = 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9$$
$$S = \varepsilon \mid 0 S \mid D S$$
$$N = 0 \mid D S$$

Contrast with “duzentos e trinta e seite”

Specialization, limited vocabulary, simple grammar

# 13<sup>th</sup> century



Specialization, limited vocabulary, simple grammar

## 16<sup>th</sup> century

$$x^3 + 3x^2 = 20$$

instead of a cube and three squares make twenty  
or a cube of the thing and three squares of the thing make twenty

Specialization, limited vocabulary, simple grammar

## 19<sup>th</sup> century

$$\forall x \exists y (x = 2 \times y \vee x = 2 \times y + 1)$$

Specialization, limited vocabulary, simple grammar

Specialization: all declarative sentences, but only declarative sentences (for instance, no insults)

## 19<sup>th</sup> century

acetylseryltyrosylserylisoleucylthreonylserylproylserylglutaminyl-  
phenylalanylvalylphenylalanylleucylserylserylvalyltryptophylalanyl-  
aspartylproylisoleucylglutamylleucylleucylasparaginyvalylcysteinyl-  
threonylserylserylleucylglycylasparaginyglutaminylphenylalanyl-  
glutaminylthreonylglutaminylglutaminylalanylarginylthreonylthreonyl-  
glutaminylvalylglutaminylglutaminylphenylalanylserylglutaminylvalyl-  
tryptophyllsylproylphenylalanylproylglutaminylserylthreonylvalyl-  
arginylphenylalanylproylglycylaspartylvalyltyrosyllsylvalyltyrosyl-  
arginyltyrosylasparaginyalanylvalylleucylaspartylproylleucylisoleucyl-

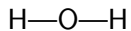
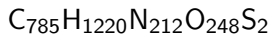
...

...

threonylalanylleucylleucylglycylthreonylphenylalanylasparylthreonyl-  
arginylasparaginyllarginylisoleucylisoleucylglutamylvalylglutamyl-  
asparaginyllglutaminyllglutaminyllserylprolylthreonylthreonylalanylglutamyl-  
threonylleucylaspartylalanylthreonyllarginylarginylvalylaspartylaspartyl-  
alanylthreonylvalylalanylisoleucylarginylseryllalanylasparginylisoleucyl-  
asparaginyllleucylvalylasparaginyllglutamylleucylvalylarginylglycyl-  
threonyllglycylleucyltyrosylasparaginyllglutaminyllasparaginyllthreonyl-  
phenylalanylglutamylseryllmethionylseryllglycylleucylvalyltryptophyl-  
threonyllseryllalanylprolyllalanylserine

chemical nomenclature

And also





## In between: the language of science

~~let poodle  $\equiv$  feijoada; I want to eat a poodle~~

We shall call *work* the scalar product of a force by the displacement of its point of application. When the displacement is orthogonal to the force, the *work* is null.

Illimited scope

Let  $x$  be  $y + 4$ , as  $y$  is even, so is  $x$

Limited scope

## In between: legal language

Lessor, Mr. Fulano, shall be referred to as "OWNER" and Lessee, Mrs. Cicrano, shall be referred to as "RESIDENT."

Limited scope

# The price of universality of natural languages

Inability to express things precisely

When go deeper in science, law, music, etc.

move away from natural language

sometimes just a little bit

sometimes dramatically: formal languages

# A common language with computers?

The language of algebra

$$x^3 + 3x^2 = 20$$

not been invented to communicate with computers

But because natural languages are too **cumbersome** to express equations

Also too cumbersome to express algorithms

Hence **programming languages**

## V. Formal languages and the invention of writing

# The victory of positional notation

No way to give a name to each number: 1, 2, 3, etc.

Decide to stop at 9 and decompose larger numbers

CC XXX IIIIIII

2C 3X 7I

237

(as well as CCXXXVII, etc.)

Only symbols for the powers of ten I, X, C, M, etc.

Only symbols for the numbers between 0 and 9

Both

Up to MMMMMMMMMM CCCCCCCCCC XXXXXXXXXXXX IIIIIIIII

Up to 9M 9C 9X 9I

All numbers

## A partial victory

Written language: 237

Spoken language: two hundred thirty seven

Two is a number of hundreds: a word “hundred” not a position

Like: 2C 3X 7I

Why not saying **two - three - seven**?

Too sensitive to perturbations

two **hundred** thirty seven

two **thousands** three hundred seventy one

Quickly get the order or magnitude, even if end of message lost

Not in **two - three - seven** and **two - three - seven - one**

# Positional notation and written language

Positional notation cannot be used in spoken language

It cannot have been invented before writing

All languages: no way to read loud *NENWNEESESSWWN*, acetylseryl..., a program, a Feynman diagram, an architectural drawing, etc.

No formal language before writing



# Formal and written languages

The invention of writing has opened possibilities to express ourselves differently **in formal languages**

May be: an **illusion** that written language has been invented to transcribe spoken language (because spoken language learned before written language)

What has writing been invented for?

# The oldest traces of writing



Transformation of clay bullae



When entrusting a flock to a shepherd

Put in a bulla as many tokens as head of cattle

Open it to check that the shepherd has not stolen an animal

Then: tokens replaced by notches on the bulla

# Bullae

All the characteristics of a formal language:

- ▶ very specialized
- ▶ limited vocabulary (tokens)
- ▶ simple grammar (put in the bulla and seal)

Writing invented to express a formal language, not a natural one

## Later

Writing language got more complex to transcribe the Sumerian spoken language

Writing seems to be a technique to express formal languages  
(partially) diverted to express natural languages

# A science of formal languages?

What does informatics has to do with logic, grammar, music, and law?

Pieces of a general science of formal languages, **yet to come**