Formal Languages applied to Linguistics

Pascal Amsili

Laboratoire Lattice, Université Sorbonne Nouvelle

Cogmaster, september 2019

Good questions

Why would one consider natural language as a formal language?

- it allows to describe the language in a formal/compact/elegant way
- it allows to compare various languages (via classes of languages established by mathematicians)
- it give algorithmic tools to recognize and to analyse words of a language.

recognize u	: decide whether $u \in L$
analyse u	: show the internal structure of u

Overview

1. Formal Languages
2. Formal Grammars
   - Definition
   - Language classes
3. Regular Languages
4. Formal complexity of Natural Languages

Introduction

Formal grammars have been proposed by Chomsky as **one of the available means** to characterize a formal language.
Other means include :

- Turing machines (automata)
- λ-terms
- ...

Formal grammar

Def. 9 ((Formal) Grammar)

A **formal grammar** is defined by $\langle \Sigma, N, S, P \rangle$ where

- Σ is an alphabet
- N is a disjoint alphabet non-terminal vocabulary)
- $S \in V$ is a distinguished elemnt of N, called the *axiom*
- P is a set of « *production rules* », namely a subset of the cartesian product $(\Sigma \cup N)^* N (\Sigma \cup N)^* \times (\Sigma \cup N)^*$.

Examples

$$\langle \Sigma, N, S, P \rangle$$

$$\mathcal{G}_0 = \Big\langle$$

Examples

$$\langle \Sigma, N, S, P \rangle$$

$$\mathcal{G}_0 = \Big\langle \{joe, sam, sleeps\},$$

Examples

$$\langle \Sigma, N, S, P \rangle$$

$$\mathcal{G}_0 = \Big\langle \{joe, sam, sleeps\}, \{N, V, S\},$$

Examples

$$\langle \Sigma, N, S, P \rangle$$

$$\mathcal{G}_0 = \Bigg\langle \{joe, sam, sleeps\}, \{N, V, S\}, S,$$

Examples

$$\langle \Sigma, N, S, P \rangle$$

$$\mathcal{G}_0 = \left\langle \{joe, sam, sleeps\}, \{N, V, S\}, S, \left\{ \begin{array}{l} (N, joe) \\ (N, sam) \\ (V, sleeps) \\ (S, N\ V) \end{array} \right\} \right\rangle \}$$

Examples

$$\langle \Sigma, N, S, P \rangle$$

$$\mathcal{G}_0 = \left\langle \{joe, sam, sleeps\}, \{N, V, S\}, S, \left\{ \begin{array}{l} N \rightarrow joe \\ N \rightarrow sam \\ V \rightarrow sleeps \\ S \rightarrow N\ V \end{array} \right\} \right\rangle \}$$

Examples (cont'd)

$$\mathcal{G}_1 = \left\langle \{jean, dort\}, \{Np, SN, SV, V, S\}, S, \left\{ \begin{array}{l} S \rightarrow SN\ SV \\ SN \rightarrow Np \\ SV \rightarrow V \\ Np \rightarrow jean \\ V \rightarrow dort \end{array} \right\} \right\rangle \}$$

$$\mathcal{G}_2 = \langle \{(,)\}, \{S\}, S, \{S \longrightarrow \varepsilon \,|\, (S)S\} \rangle$$

Notation

$$
\begin{array}{llll}
\mathcal{G}_3: & E & \longrightarrow & E + E \\
 & & | & E \times E \\
 & & | & (\ E\ ) \\
 & & | & F \\
 & F & \longrightarrow & 0 \,|\, 1 \,|\, 2 \,|\, 3 \,|\, 4 \,|\, 5 \,|\, 6 \,|\, 7 \,|\, 8 \,|\, 9
\end{array}
$$

Notation

$$
\begin{array}{llll}
\mathcal{G}_3 : & E & \longrightarrow & E + E \\
& & | & E \times E \\
& & | & (\ E\ ) \\
& & | & F \\
& F & \longrightarrow & 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9
\end{array}
$$

$$\mathcal{G}_3 = \langle \{+, \times, (,), 0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}, \{E, F\}, E, \{\ldots\} \rangle$$

Notation

$$
\begin{array}{llll}
\mathcal{G}_3: & E & \longrightarrow & E + E \\
 & & | & E \times E \\
 & & | & (\ E\ ) \\
 & & | & F \\
 & F & \longrightarrow & 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9
\end{array}
$$

$$\mathcal{G}_3 = \langle \{+, \times, (,), 0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}, \{E, F\}, E, \{\ldots\} \rangle$$

$$G_4 = E \rightarrow E + T \mid T, T \rightarrow T \times F \mid F, F \rightarrow (E) \mid a$$

Immediate Derivation

Def. 10 (Immediate derivation)

Let $\mathcal{G} = \langle X, V, S, P \rangle$ a grammar, $(f, g) \in (X \cup V)^*$ two "words", $r \in P$ a production rule, such that $r : A \longrightarrow u$ ($u \in (X \cup V)^*$).

- f derives into g (immediate derivation) with the rule r (noted $f \xrightarrow{r} g$) iff
  $\exists v, w$ s.t. $f = vAw$ and $g = vuw$
- f derives into g (immediate derivation) in the grammar $\mathcal{G}$ (noted $f \xrightarrow{\mathcal{G}} g$) iff
  $\exists r \in P$ s.t. $f \xrightarrow{r} g$.

Derivation

Def. 11 (Derivation)

$$f \xrightarrow{\mathcal{G}*} g \text{ if } \begin{cases} f = g & \text{or} \\ \exists f_0, f_1, f_2, ..., f_n \text{ s.t. } & f_0 = f \\ & f_n = g \\ & \forall i \in [1, n] : f_{i-1} \xrightarrow{\mathcal{G}} f_i \end{cases}$$

An example with $\mathcal{G}_0$:
N V joe N

Derivation

Def. 11 (Derivation)

$$f \xrightarrow{\mathcal{G}*} g \text{ if } \begin{cases} f = g & \text{or} \\ \exists f_0, f_1, f_2, \ldots, f_n \text{ s.t. } & f_0 = f \\ & f_n = g \\ & \forall i \in [1, n] : f_{i-1} \xrightarrow{\mathcal{G}} f_i \end{cases}$$

An example with $\mathcal{G}_0$:

$N\ V\ joe\ N \longrightarrow sam\ V\ joe\ N$

Derivation

Def. 11 (Derivation)

$f \xrightarrow{\mathcal{G}*} g$ if $f = g$ or

$\exists f_0, f_1, f_2, \ldots, f_n$ s.t. $f_0 = f$

$f_n = g$

$\forall i \in [1, n] : f_{i-1} \xrightarrow{\mathcal{G}} f_i$

An example with $\mathcal{G}_0$:

$N\ V\ joe\ N \longrightarrow sam\ V\ joe\ N \longrightarrow sam\ V\ joe\ joe$ or

Derivation

Def. 11 (Derivation)

$$f \xrightarrow{\mathcal{G}*} g \text{ if } \begin{array}{ll} f = g & \text{or} \\ \exists f_0, f_1, f_2, ..., f_n \text{ s.t. } \begin{array}{l} f_0 = f \\ f_n = g \\ \forall i \in [1, n] : f_{i-1} \xrightarrow{\mathcal{G}} f_i \end{array} & \end{array}$$

An example with $\mathcal{G}_0$:

$$N\ V\ joe\ N \longrightarrow sam\ V\ joe\ N \longrightarrow \begin{array}{ll} sam\ V\ joe\ joe & \text{or} \\ sam\ V\ joe\ sam & \text{or} \end{array}$$

Derivation

Def. 11 (Derivation)

$$f \xrightarrow{\mathcal{G}*} g \text{ if } \begin{cases} f = g & \text{or} \\ \exists f_0, f_1, f_2, \dots, f_n \text{ s.t. } & f_0 = f \\ & f_n = g \\ & \forall i \in [1, n] : f_{i-1} \xrightarrow{\mathcal{G}} f_i \end{cases}$$

An example with $\mathcal{G}_0$:

$$N\ V\ joe\ N \longrightarrow sam\ V\ joe\ N \longrightarrow \begin{array}{ll} sam\ V\ joe\ joe & \text{or} \\ sam\ V\ joe\ sam & \text{or} \\ sam\ sleeps\ joe\ N & \text{or} \\ \dots & \end{array}$$

Endpoint of a derivation

$$
\begin{array}{llll}
\mathcal{G}_3: & E & \longrightarrow & E + E \\
 & & | & E \times E \\
 & & | & (\ E\ ) \\
 & & | & F \\
 & F & \longrightarrow & 0 \,|\, 1 \,|\, 2 \,|\, 3 \,|\, 4 \,|\, 5 \,|\, 6 \,|\, 7 \,|\, 8 \,|\, 9
\end{array}
$$

An example with $\mathcal{G}_3$:

$E \times E$

Endpoint of a derivation

$$\begin{array}{llll} \mathcal{G}_3: & E & \longrightarrow & E + E \\ & & | & E \times E \\ & & | & (\ E\ ) \\ & & | & F \\ & F & \longrightarrow & 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9 \end{array}$$

An example with $\mathcal{G}_3$:

$$E \times E \longrightarrow F \times E$$

Endpoint of a derivation

$$
\begin{array}{llll}
\mathcal{G}_3 : & E & \longrightarrow & E + E \\
& & | & E \times E \\
& & | & (E) \\
& & | & F \\
& F & \longrightarrow & 0 \,|\, 1 \,|\, 2 \,|\, 3 \,|\, 4 \,|\, 5 \,|\, 6 \,|\, 7 \,|\, 8 \,|\, 9
\end{array}
$$

An example with $\mathcal{G}_3$:

$$E \times E \longrightarrow F \times E \longrightarrow 3 \times E$$

Endpoint of a derivation

$$\begin{array}{llll} \mathcal{G}_3: & E & \longrightarrow & E + E \\ & & | & E \times E \\ & & | & (\ E\ ) \\ & & | & F \\ & F & \longrightarrow & 0 \,|\, 1 \,|\, 2 \,|\, 3 \,|\, 4 \,|\, 5 \,|\, 6 \,|\, 7 \,|\, 8 \,|\, 9 \end{array}$$

An example with $\mathcal{G}_3$:

$$E \times E \longrightarrow F \times E \longrightarrow 3 \times E \longrightarrow 3 \times (E)$$

Endpoint of a derivation

$$
\begin{array}{llll}
\mathcal{G}_3 : & E & \longrightarrow & E + E \\
& & | & E \times E \\
& & | & (\ E\ ) \\
& & | & F \\
& F & \longrightarrow & 0 \,|\, 1 \,|\, 2 \,|\, 3 \,|\, 4 \,|\, 5 \,|\, 6 \,|\, 7 \,|\, 8 \,|\, 9
\end{array}
$$

An example with $\mathcal{G}_3$:

$$E \times E \longrightarrow F \times E \longrightarrow 3 \times E \longrightarrow 3 \times (E) \longrightarrow 3 \times (E + E)$$

Endpoint of a derivation

$$
\begin{array}{llll}
\mathcal{G}_3: & E & \longrightarrow & E + E \\
 & & | & E \times E \\
 & & | & (\ E\ ) \\
 & & | & F \\
 & F & \longrightarrow & 0 \,|\, 1 \,|\, 2 \,|\, 3 \,|\, 4 \,|\, 5 \,|\, 6 \,|\, 7 \,|\, 8 \,|\, 9
\end{array}
$$

An example with $\mathcal{G}_3$:

$$E \times E \longrightarrow F \times E \longrightarrow 3 \times E \longrightarrow 3 \times (E) \longrightarrow 3 \times (E + E) \longrightarrow 3 \times (E + F)$$

Endpoint of a derivation

$$
\begin{array}{llll}
\mathcal{G}_3 : & E & \longrightarrow & E + E \\
& & | & E \times E \\
& & | & (\ E\ ) \\
& & | & F \\
& F & \longrightarrow & 0\,|\,1\,|\,2\,|\,3\,|\,4\,|\,5\,|\,6\,|\,7\,|\,8\,|\,9
\end{array}
$$

An example with $\mathcal{G}_3$:

$$E \times E \longrightarrow F \times E \longrightarrow 3 \times E \longrightarrow 3 \times (E) \longrightarrow 3 \times (E + E) \longrightarrow 3 \times (E + F) \longrightarrow 3 \times (E + 4)$$

Endpoint of a derivation

$$\mathcal{G}_3 : \begin{array}{lcl} E & \longrightarrow & E + E \\ & | & E \times E \\ & | & (\ E\ ) \\ & | & F \\ F & \longrightarrow & 0\,|\,1\,|\,2\,|\,3\,|\,4\,|\,5\,|\,6\,|\,7\,|\,8\,|\,9 \end{array}$$

An example with $\mathcal{G}_3$:

$$E \times E \longrightarrow F \times E \longrightarrow 3 \times E \longrightarrow 3 \times (E) \longrightarrow 3 \times (E + E) \longrightarrow$$
$$3 \times (E + F) \longrightarrow 3 \times (E + 4) \longrightarrow 3 \times (F + 4)$$

Endpoint of a derivation

$$
\begin{array}{llll}
\mathcal{G}_3 : & E & \longrightarrow & E + E \\
 & & | & E \times E \\
 & & | & (\; E \;) \\
 & & | & F \\
 & F & \longrightarrow & 0 \,|\, 1 \,|\, 2 \,|\, 3 \,|\, 4 \,|\, 5 \,|\, 6 \,|\, 7 \,|\, 8 \,|\, 9
\end{array}
$$

An example with $\mathcal{G}_3$:

$$
E \times E \longrightarrow F \times E \longrightarrow 3 \times E \longrightarrow 3 \times (E) \longrightarrow 3 \times (E + E) \longrightarrow \\
3 \times (E + F) \longrightarrow 3 \times (E + 4) \longrightarrow 3 \times (F + 4) \longrightarrow 3 \times (5 + 4)
$$

Endpoint of a derivation

$$
\begin{array}{llll}
\mathcal{G}_3: & E & \longrightarrow & E + E \\
 & & | & E \times E \\
 & & | & (\; E \;) \\
 & & | & F \\
 & F & \longrightarrow & 0 \,|\, 1 \,|\, 2 \,|\, 3 \,|\, 4 \,|\, 5 \,|\, 6 \,|\, 7 \,|\, 8 \,|\, 9
\end{array}
$$

An example with $\mathcal{G}_3$:

$$
E \times E \longrightarrow F \times E \longrightarrow 3 \times E \longrightarrow 3 \times (E) \longrightarrow 3 \times (E + E) \longrightarrow \\
3 \times (E + F) \longrightarrow 3 \times (E + 4) \longrightarrow 3 \times (F + 4) \longrightarrow 3 \times (5 + 4) \longrightarrow\!\!|
$$

Engendered language

Def. 12 (Language engendered by a word)

Let $f \in (\Sigma \cup N)^*$.

$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$

Def. 13 (Language engendered by a grammar)

The *language engendered by a grammar* $\mathcal{G}$ is the set of words of Σ^* derived from the axiom.

$L_{\mathcal{G}} = L_{\mathcal{G}}(S)$

Engendered language

Def. 12 (Language engendered by a word)

Let $f \in (\Sigma \cup N)^*$.

$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$

Def. 13 (Language engendered by a grammar)

The *language engendered by a grammar* $\mathcal{G}$ is the set of words of Σ^* derived from the axiom.

$L_{\mathcal{G}} = L_{\mathcal{G}}(S)$

For instance $() \in L_{\mathcal{G}_2}$:

Engendered language

Def. 12 (Language engendered by a word)

Let $f \in (\Sigma \cup N)^*$.

$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$

Def. 13 (Language engendered by a grammar)

The *language engendered by a grammar* $\mathcal{G}$ is the set of words of Σ^* derived from the axiom.

$L_{\mathcal{G}} = L_{\mathcal{G}}(S)$

For instance $() \in L_{\mathcal{G}_2}$: $S \rightarrow (S)S$

Engendered language

Def. 12 (Language engendered by a word)

Let $f \in (\Sigma \cup N)^*$.

$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$

Def. 13 (Language engendered by a grammar)

The *language engendered by a grammar* $\mathcal{G}$ is the set of words of Σ^* derived from the axiom.

$L_{\mathcal{G}} = L_{\mathcal{G}}(S)$

For instance $() \in L_{\mathcal{G}_2}$: $S \rightarrow (S)S \rightarrow ()S$

Engendered language

Def. 12 (Language engendered by a word)

Let $f \in (\Sigma \cup N)^*$.

$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$

Def. 13 (Language engendered by a grammar)

The *language engendered by a grammar* $\mathcal{G}$ is the set of words of Σ^* derived from the axiom.

$L_{\mathcal{G}} = L_{\mathcal{G}}(S)$

For instance $() \in L_{\mathcal{G}_2}$: $S \rightarrow (S)S \rightarrow ()S \rightarrow ()$

Engendered language

Def. 12 (Language engendered by a word)

Let $f \in (\Sigma \cup N)^*$.

$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$

Def. 13 (Language engendered by a grammar)

The *language engendered by a grammar* $\mathcal{G}$ is the set of words of Σ^* derived from the axiom.

$L_{\mathcal{G}} = L_{\mathcal{G}}(S)$

For instance $() \in L_{\mathcal{G}_2}$: $S \rightarrow (S)S \rightarrow ()S \rightarrow ()$
as well as $((()))$, $()()()$, $((()()()))\ldots$

Engendered language

Def. 12 (Language engendered by a word)

Let $f \in (\Sigma \cup N)^*$.

$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$

Def. 13 (Language engendered by a grammar)

The *language engendered by a grammar* $\mathcal{G}$ is the set of words of Σ^* derived from the axiom.

$L_{\mathcal{G}} = L_{\mathcal{G}}(S)$

For instance $() \in L_{\mathcal{G}_2}$: $S \rightarrow (S)S \rightarrow ()S \rightarrow ()$
as well as $((()))$, $()()()$, $((()()()))\ldots$
but $)()(\notin L_{\mathcal{G}_2}$, even though the following is a licit derivation :

Engendered language

Def. 12 (Language engendered by a word)

Let $f \in (\Sigma \cup N)^*$.

$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$

Def. 13 (Language engendered by a grammar)

The *language engendered by a grammar* $\mathcal{G}$ is the set of words of Σ^* derived from the axiom.

$L_{\mathcal{G}} = L_{\mathcal{G}}(S)$

For instance $() \in L_{\mathcal{G}_2}$: $S \rightarrow (S)S \rightarrow ()S \rightarrow ()$
as well as $((()))$, $()()()$, $((()()()))\ldots$
but $)()(\notin L_{\mathcal{G}_2}$, even though the following is a licit derivation :
$)S(\rightarrow$

Engendered language

Def. 12 (Language engendered by a word)

Let $f \in (\Sigma \cup N)^*$.

$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$

Def. 13 (Language engendered by a grammar)

The *language engendered by a grammar* $\mathcal{G}$ is the set of words of Σ^* derived from the axiom.

$L_{\mathcal{G}} = L_{\mathcal{G}}(S)$

For instance $() \in L_{\mathcal{G}_2}$: $S \rightarrow (S)S \rightarrow ()S \rightarrow ()$
as well as $((()))$, $()()()$, $((()()()))\ldots$
but $)()(\notin L_{\mathcal{G}_2}$, even though the following is a licit derivation :
$)S(\rightarrow)(S)S(\rightarrow$

Engendered language

Def. 12 (Language engendered by a word)

Let $f \in (\Sigma \cup N)^*$.
$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$

Def. 13 (Language engendered by a grammar)

The *language engendered by a grammar* $\mathcal{G}$ is the set of words of Σ^* derived from the axiom.
$L_{\mathcal{G}} = L_{\mathcal{G}}(S)$

For instance $() \in L_{\mathcal{G}_2}$: $S \rightarrow (S)S \rightarrow ()S \rightarrow ()$
as well as $((()))$, $()()()$, $((()()()))$...
but $)()(\notin L_{\mathcal{G}_2}$, even though the following is a licit derivation :
$)S(\rightarrow)(S)S(\rightarrow)()S(\rightarrow$

Engendered language

Def. 12 (Language engendered by a word)

Let $f \in (\Sigma \cup N)^*$.

$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$

Def. 13 (Language engendered by a grammar)

The *language engendered by a grammar* $\mathcal{G}$ is the set of words of Σ^* derived from the axiom.

$L_{\mathcal{G}} = L_{\mathcal{G}}(S)$

For instance $() \in L_{\mathcal{G}_2}$: $S \rightarrow (S)S \rightarrow ()S \rightarrow ()$
as well as $((()))$, $()()()$, $((()()()))\ldots$
but $)()(\notin L_{\mathcal{G}_2}$, even though the following is a licit derivation :
$)S(\rightarrow)(S)S(\rightarrow)()S(\rightarrow)()($

Engendered language

Def. 12 (Language engendered by a word)

Let $f \in (\Sigma \cup N)^*$.

$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$

Def. 13 (Language engendered by a grammar)

The *language engendered by a grammar* $\mathcal{G}$ is the set of words of Σ^* derived from the axiom.

$L_{\mathcal{G}} = L_{\mathcal{G}}(S)$

For instance $() \in L_{\mathcal{G}_2}$: $S \rightarrow (S)S \rightarrow ()S \rightarrow ()$
as well as $((()))$, $()()()$, $((()()()))\ldots$
but $)()(\notin L_{\mathcal{G}_2}$, even though the following is a licit derivation :
$)S(\rightarrow)(S)S(\rightarrow)()S(\rightarrow)()($
for there is no way to arrive at $)S($ starting with S.

Example

$$G_4 = E \to E + T \mid T, T \to T \times F \mid F, F \to (E) \mid a$$

$a + a$, $a + (a \times a)$, ...

Proto-word

Def. 14 (Proto-word)

A proto-word (or proto-sentence) is a word on $(\Sigma \cup N)^* N (\Sigma \cup N)^*$ (that is, a word containing at least one letter of N) produced by a derivation from the axiom.

$$E \rightarrow E + T \rightarrow E + T * F \rightarrow T + T * F \rightarrow T + F * F \rightarrow T + a * F \rightarrow F + a * F \rightarrow a + a * F \rightarrow \cancel{a + a * a}$$

Multiple derivations

A given word may have several derivations:

$E \rightarrow E + E \rightarrow F + E \rightarrow F + F \rightarrow 3 + F \rightarrow 3 + 4$

Multiple derivations

A given word may have several derivations:

$E \rightarrow E + E \rightarrow F + E \rightarrow F + F \rightarrow 3 + F \rightarrow 3 + 4$

$E \rightarrow E + E \rightarrow E + F \rightarrow E + 4 \rightarrow F + 4 \rightarrow 3 + 4$

Multiple derivations

A given word may have several derivations:

$E \to E + E \to F + E \to F + F \to 3 + F \to 3 + 4$

$E \to E + E \to E + F \to E + 4 \to F + 4 \to 3 + 4$

... but if the grammar is not ambiguous, there is only one **left** derivation:

Multiple derivations

A given word may have several derivations:

$E \rightarrow E + E \rightarrow F + E \rightarrow F + F \rightarrow 3 + F \rightarrow 3 + 4$

$E \rightarrow E + E \rightarrow E + F \rightarrow E + 4 \rightarrow F + 4 \rightarrow 3 + 4$

... but if the grammar is not ambiguous, there is only one **left** derivation:

$\underline{E} \rightarrow \underline{E} + E \rightarrow \underline{F} + E \rightarrow 3 + \underline{E} \rightarrow 3 + \underline{F} \rightarrow 3 + 4$

Multiple derivations

A given word may have several derivations:

$E \rightarrow E + E \rightarrow F + E \rightarrow F + F \rightarrow 3 + F \rightarrow 3 + 4$

$E \rightarrow E + E \rightarrow E + F \rightarrow E + 4 \rightarrow F + 4 \rightarrow 3 + 4$

... but if the grammar is not ambiguous, there is only one **left** derivation:

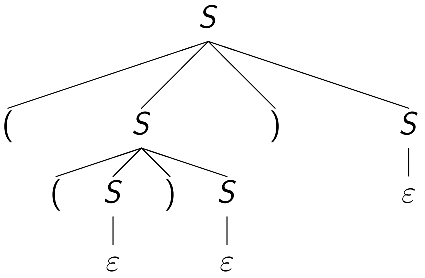
$\underline{E} \rightarrow \underline{E} + E \rightarrow \underline{F} + E \rightarrow 3 + \underline{E} \rightarrow 3 + \underline{F} \rightarrow 3 + 4$

parsing: trying to find the/a left derivation (resp. right)

Derivation tree

For context-free languages, there is a way to represent the set of equivalent derivations, via a derivation tree which shows all the derivation independantly of their order.

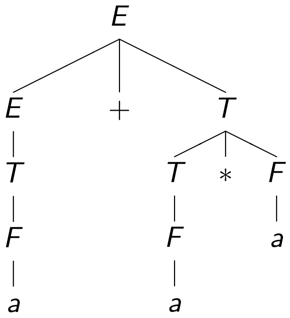
Grammar $\mathcal{G}_2$: $S \longrightarrow \varepsilon \mid (S)S$



$S \rightarrow (S)S \rightarrow ((S)S)S \rightarrow ((S)S) \rightarrow ((S)) \rightarrow (())$

Structural analysis

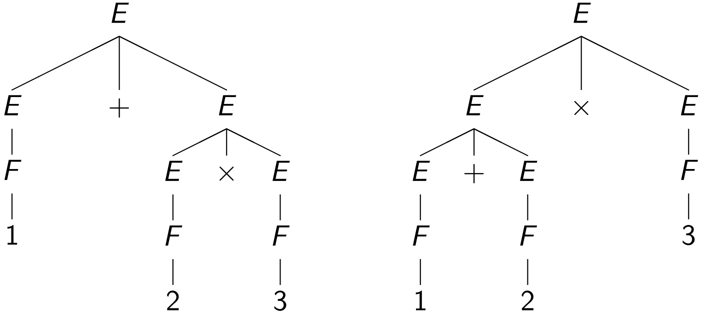
Syntactic trees are precious to give access to the semantics



Ambiguity

When a grammar can assign more than one derivation tree to a word $w \in L(G)$ (or more than one left derivation), the grammar is *ambiguous*.

For instance, $\mathcal{G}_3$ is ambiguous, since it can assign the two follwing trees to $1 + 2 \times 3$:



About ambiguity

- Ambiguity is not desirable for the semantics
- Useful artificial languages are rarely ambiguous
- There are context-free languages that are intrinsequely ambiguous (3)
- Natural languages are notoriously ambiguous...

(3) $\{a^n b a^m b a^p b a^q | (n \geqslant q \wedge m \geqslant p) \vee (n \geqslant m \wedge p \geqslant q)\}$

Comparison of grammars

- different languages generated ⇒ different grammars
- same language generated by $\mathcal{G}$ and $\mathcal{G}'$:
  ⇒ same weak generative power
- same language generated by $\mathcal{G}$ and $\mathcal{G}'$, and same structural decomposition : ⇒ same strong generative power

Overview

1. Formal Languages
2. Formal Grammars
   - Definition
   - Language classes
3. Regular Languages
4. Formal complexity of Natural Languages

Principle

Define language families on the basis of properties of the grammars that generate them :

1. Four classes are defined, they are included one in another
2. A language is of type k if it **can** be recognized by a type k grammar (and thus, by definition, by a type $k-1$ grammar) ; and **cannot** be recognized by a grammar of type $k+1$.

Chomsky's hierarchy

type 0 No restriction on $P \subset (X \cup V)^* V (X \cup V)^* \times (X \cup V)^*$.

type 1 (*context-sensitive* grammars) All rules of P are of the shape $(u_1 S u_2, u_1 m u_2)$, where u_1 and $u_2 \in (X \cup V)^*$, $S \in V$ and $m \in (X \cup V)^+$.

type 2 (*context-free* grammar) All rules of P are of the shape (S, m), where $S \in V$ and $m \in (X \cup V)^*$.

type 3 (*regular* grammars) All rules of P are of the shape (S, m), where $S \in V$ and $m \in X.V \cup X \cup \{\varepsilon\}$.

Examples

type 3:

$$
\begin{array}{lcl}
S & \rightarrow & aS \mid aB \mid bB \mid cA \\
B & \rightarrow & bB \mid b \\
A & \rightarrow & cS \mid bB
\end{array}
$$

Examples

type 3:

$$\begin{array}{lcl} S & \rightarrow & aS \mid aB \mid bB \mid cA \\ B & \rightarrow & bB \mid b \\ A & \rightarrow & cS \mid bB \end{array}$$

type 2:

$$E \rightarrow E + T \mid T, T \rightarrow T \times F \mid F, F \rightarrow (E) \mid a$$

Example 1 type 0

Type 0:

$$
\begin{array}{lll}
S \rightarrow SABC & AC \rightarrow CA & A \rightarrow a \\
S \rightarrow \varepsilon & CA \rightarrow AC & B \rightarrow b \\
AB \rightarrow BA & BC \rightarrow CB & C \rightarrow c \\
BA \rightarrow AB & CB \rightarrow BC &
\end{array}
$$

generated language :

Example 1 type 0

Type 0:

$$\begin{array}{llll} S & \rightarrow SABC & AC \rightarrow CA & A \rightarrow a \\ S & \rightarrow \varepsilon & CA \rightarrow AC & B \rightarrow b \\ AB & \rightarrow BA & BC \rightarrow CB & C \rightarrow c \\ BA & \rightarrow AB & CB \rightarrow BC & \end{array}$$

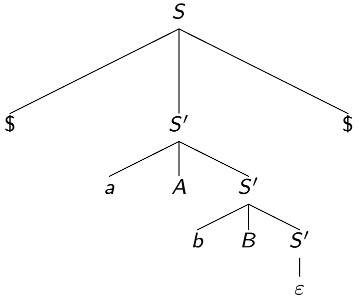
generated language : words with an equal number of a, b, and c.

Example 2: type 0

Type 0:

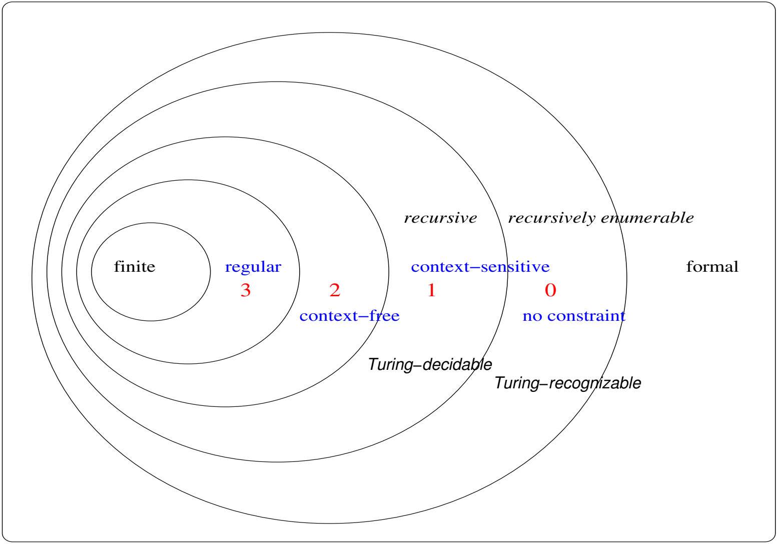
$$
\begin{array}{lll}
S \rightarrow \$S'\$ & Aa \rightarrow aA & \$a \rightarrow a\$ \\
S' \rightarrow aAS' & Ab \rightarrow bA & \$b \rightarrow b\$ \\
S' \rightarrow bBS' & Ba \rightarrow aB & A\$ \rightarrow \$a \\
S' \rightarrow \varepsilon & Bb \rightarrow bB & B\$ \rightarrow \$b \\
 & & \$\$ \rightarrow \# \\
\end{array}
$$

Example 2: type 0 (cont'd)



$\$$	a	A	b	B	$\$$
a	$\$$	A	b	B	$\$$
a	$\$$	A	b	$\$$	b
a	$\$$	b	A	$\$$	b
a	b	$\$$	A	$\$$	b
a	b	$\$$	$\$$	a	b
a	b	$\#$		a	b

Language families



Remarks

- There are others ways to classify languages,
  - either on other properties of the grammars;
  - or on other properties of the languages
- Nested structures are preferred, but it's not necessary
- When classes are nested, it is expected to have a growth of complexity/expressive power

Taking stock

What we've seen so far

- alphabet, word, concatenation, language
- operations on languages : $\cup$, ., * ...
- formal grammars : rewriting devices
- classes of grammars/languages/problems

Today's programme:

- play with a couple of grammars
- a word about syntax
- main topic: regular languages and automata

Let's play with grammars

For each of the following grammars, give the generated language, and the type they have in Chomsky's hierarchy.

$$
\begin{aligned}
S &\rightarrow S_1 S_2 \\
S_1 &\rightarrow a S_1 b \mid ab \\
S_2 &\rightarrow c S_2 \mid c
\end{aligned}
$$

$$
\begin{aligned}
S &\rightarrow aSBC \\
S &\rightarrow aBC \\
CB &\rightarrow BC \\
aB &\rightarrow ab \\
bB &\rightarrow bb \\
bC &\rightarrow bc \\
cC &\rightarrow cc
\end{aligned}
$$

Let's play with grammars (cont'd)

Give a contex-free grammar that generates each of the following languages (alphabet $\Sigma = \{a, b, c\}$).

- $L_0 = \{w \in X^* \ / \ w = a^n \ ; \ n \geq 0\}$
- $L'_0 = \{w \in X^* \ / \ w = a^n b^n ca \ ; \ n \geq 0\}$
- $L_1 = \{w \in X^* \ / \ w = a^n b^n c^p ; n > 0 \text{ et } p > 0\}$
- $L_2 = \{w \in X^* \ / \ w = a^n b^n a^m b^m ; n, m \geq 1\}$
- $L'_3 = \{w \in X^* \ / \ |w|_a = |w]_b\}$
- $L_3 = \{w \in X^* \ / \ |w|_a = 2|w]_b\}$
- $L_4 = \{w \in X^* \ / \ \exists x \in X^* \text{ tq } w = x\overline{x}\}$
- $L_5 = \{w \in X^* \ / \ w = \overline{w}\}$

What about artificial languages? I

(i) If A is a predicate name from L_p vocabulary, and each of $t_1 ... t_n$ are constants or variables from L_p vocabulary, then $A(t_1, ..., t_n)$ is a well-formed formula (wff).

(ii) If φ is a wff, then so is $\neg\varphi$.

(iii) If φ and ψ are wffs, then $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$, $(\varphi \rightarrow \psi)$, $(\varphi \leftrightarrow \psi)$ are wffs.

(iv) If φ is a wff and x a variable, then $\forall x\varphi$ and $\exists x\varphi$ are wfss.

(v) Nothing else is a well-formed formula.

What about artificial languages? II

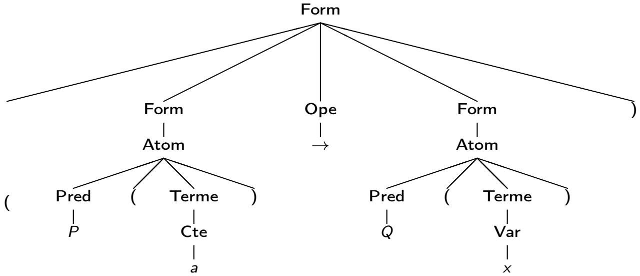
1. Terminal alphabet :
$\{\underbrace{x, y, z}_{\text{var.}}, \underbrace{a, b, c}_{\text{const.}}, \underbrace{P, Q, A, B, F}_{\text{prédicats}}, \underbrace{\wedge, \vee, \rightarrow, \leftrightarrow, \neg}_{\text{opér.}} \underbrace{(,)}_{\text{par.}}, \underbrace{\forall, \exists}_{\text{quant.}}\}$
non terminal alphabet: {Var, Cte, Pred, Terme, Quant, Ope, Atom, Form}.

$$\begin{array}{lll} \text{Var} & \rightarrow & x \mid y \mid z \\ \text{Cte} & \rightarrow & a \mid b \mid c \\ \text{Terme} & \rightarrow & \text{Var} \mid \text{Cte} \\ \text{Pred} & \rightarrow & P \mid Q \mid A \mid B \mid F \\ \text{Ope} & \rightarrow & \wedge \mid \vee \mid \rightarrow \mid \leftrightarrow \\ \text{Quant} & \rightarrow & \forall \mid \exists \end{array}$$

What about artificial languages? III

Atom	→	Pred (Terme)	
Form	→	Atom	règle (i)
	|	¬ Form	règle (ii)
	|	(Form Ope Form)	règles (iii)
	|	Quant Var Form	règles (iv)

2



References I

Bar-Hillel, Yehoshua, Perles, Micha, & Shamir, Eliahu. 1961. On formal properties of simple phrase structure grammars. *STUF-Language Typology and Universals*, 14(1-4), 143–172.

Bresnan, Joan (ed). 1982. *The Mental Representation of Grammatical Relations*. MIT Press.

Chomsky, Noam. 1957. *Syntactic Structures*. Den Haag: Mouton & Co.

Gazdar, Gerald, & Pullum, Geoffrey K. 1985 (May). *Computationally Relevant Properties of Natural Languages and Their Grammars*. Tech. rept. Center for the Study of Language and Information, Leland Stanford Junior University.

Joshi, Aravind K. 1985. *Tree Adjoining Grammars: How Much Context-Sensitivity is Required to Provide Reasonable Structural Descriptions?* Tech. rept. Department of Computer and Information Science, University of Pennsylvania.

Langendoen, D Terence, & Postal, Paul Martin. 1984. *The vastness of natural languages*. Basil Blackwell Oxford.

Mannell, Robert. 1999. *Infinite number of sentences*. part of a set of class notes on the Internet. `http://clas.mq.edu.au/speech/infinite_sentences/`.

Pollard, Carl, & Sag, Ivan A. 1994. *Head-Driven Phrase Structure Grammar*. Stanford: CSLI.

Schieber, Stuart M. 1985. Evidence against the Context-Freeness of Natural Language. *Linguistics and Philosophy*, 8(3), 333–343.

Stabler, Edward P. 2011. Computational perspectives on minimalism. *Oxford handbook of linguistic minimalism*, 617–643.

Steedman, Mark. 1988. Combinators and Grammars. *Pages 417–442 of:* Oehrle, Richard T., Bach, Emmon, & Wheeler, Deirdre (eds), *Categorical Grammars and Natural Language Structures*, vol. 32. D. Reidel Publishing Co.

Tesnière, Lucien. 1959. *Eléments de syntaxe structurale*. Librairie C. Klincksieck.