

Les pingouins shadoks

Année 2009 – 2010

Mélanie Billès, Jean Rax, Fabrice Vaillant, élèves de première S
Thomas Williams, Maximilien Burq, Emmanuel Lecouturier, élèves de terminale S

Professeurs : D. Juliot, D. Missenard

Établissement : Lycée Blaise Pascal, Orsay

Chercheurs : Pierre Pansu, Jean-Benoît Bost et Nicolas Burq, professeurs à l'Université d'Orsay Paris XI et Antoine Perasso, moniteur doctorant de Paris XI.

Introduction

Les pingouins shadoks ne savent prononcer que trois syllabes : **Ga**, **Bu** et **Meu**. Le grammairien Gaga a instauré une règle de langage : il est interdit de répéter le même groupe de syllabes consécutivement. Par exemple **Gabumeumeu** est interdit car la syllabe « **meu** » est répétée consécutivement, de même pour **Meubugabumeugabumeubu** car le groupe de syllabes « **gabumeu** » est répété consécutivement. Au contraire un mot comme **Gabumeugameubumeugabumeubuga** est correct car il n'y a aucune répétition de groupe de syllabes consécutivement. Donc si X, Y et Z sont 3 groupes de syllabes différents, un mot est interdit si il peut s'écrire XYZ avec X et Z pouvant être nuls. Pour simplifier on notera **A**, **B** et **C** les 3 syllabes différentes (cela ne veut pas dire que ga = A, bu = B et meu = C).

Les problèmes

Existe-t-il des mots de 100, 1 000 ou 100 000 syllabes ? Combien y a-t-il de mots de 100 syllabes ? Quel est le nombre minimum de syllabes pour générer une langue de 100 000 mots comme le français ? Nous avons donc essayé de trouver des très grands mots corrects et de trouver tous les mots corrects d'une longueur donnée manuellement. Cela nous a vite paru très long et fastidieux et nous avons donc essayé une approche informatique.

Comment trouver un mot de longueur n donnée ?

Particularité de certains mots

Certains mots sont finis c'est à dire que l'on ne peut leur rajouter aucune syllabe à droite. Par exemple ABACABA est fini car on ne peut rajouter ni de A car la séquence «AA» est interdite, ni de B car la séquence «ABAB» est interdite, ni de C car la séquence «ABACABAC» est interdite (Mais les mots finis sont quand même corrects).

Méthode

- 1) On part d'un mot de longueur 1.
- 2) On essaye de rajouter une syllabe.
- 3) Si le mot ainsi formé est correct alors on ajoute de nouveau une lettre.
- 4) Si le mot n'est pas correct on recommence en essayant de rajouter une autre syllabe.
- 5) Au bout de 3 fois (1), si le mot est fini on enlève la dernière lettre et on recommence. Ainsi de suite jusqu'à atteindre la taille voulue. (2)

Vérification

On veut vérifier un mot de longueur n sachant que le mot de longueur $n-1$ est correct. Donc si il y a une répétition elle inclut forcément la dernière lettre. On vérifie donc par blocs successifs jusqu'à ce que la taille des blocs soit égale à la partie entière de la moitié de la taille du mot.

Voici l'algorithme :

- Soit $\text{mot}[]$ un mot de longueur $(n-1)$ correct. $\text{Mot}[i]$ désigne la i ème lettre de mot.
- On ajoute une lettre à $\text{mot}[]$
- $j = 0$
- Tant que $j \leq \lfloor n/2 \rfloor$ (ou $\lfloor (n-1)/2 \rfloor$ si n est impair)
- On vérifie si la séquence $\text{mot}[n] \text{ mot}[n-1] \dots \text{mot}[n-j]$ n'est pas identique à la séquence $\text{mot}[n-j-1] \text{ mot}[n-j-2] \dots \text{mot}[n-2j]$
- Si les séquences sont identiques, c'est que le mot formé est interdit
- Sinon, on augmente j de 1

Exemple :

On part du mot ABACAB qui est correct.

On essaye de rajouter un A ce qui donne ABACABA.

On vérifie d'abord les deux blocs de taille 1 qui contiennent ce A.

Comme $B \neq A$ on augmente la taille du bloc. $CA \neq BA$. $BAC \neq ABA$.

On n'essaie pas avec des blocs de 4 car il n'y en a qu'un seul qui contient la dernière lettre.

Le mot est donc correct.

On essaye de rajouter un A ce qui donne le mot ABACABAA. Or $A = A$. Donc le mot présente une répétition, il est donc interdit.

On essaye de rajouter un B ce qui donne ABACABAB. $A \neq B$. Mais $AB = AB$. Donc le mot est interdit.

On essaye de rajouter un C ce qui donne ABACABAC. $A \neq C$. $AB \neq AC$. $ACA \neq BAC$. Mais $ABAC = ABAC$ donc ce mot est interdit.

On ne peut donc rajouter ni de A ni de B ni de C à ABACABA qui est par conséquent fini.

On enlève donc le A final et on essaye de mettre un B à la place ce qui donne ABACABB. $B = B$ donc le mot est interdit. On essaye de rajouter un C ce qui donne ABACABC. $B \neq C$. $CA \neq BC$. $BAC \neq ABC$ donc le mot est correct.

On essaye donc de rajouter à ABACABC un A ce qui donne ABACABCA. $C \neq A$. $AB \neq CA$. $ACA \neq BCA$.

ABAC $\neq$ ABCA donc le mot est correct.

On continue ainsi jusqu'à obtenir la taille voulue.

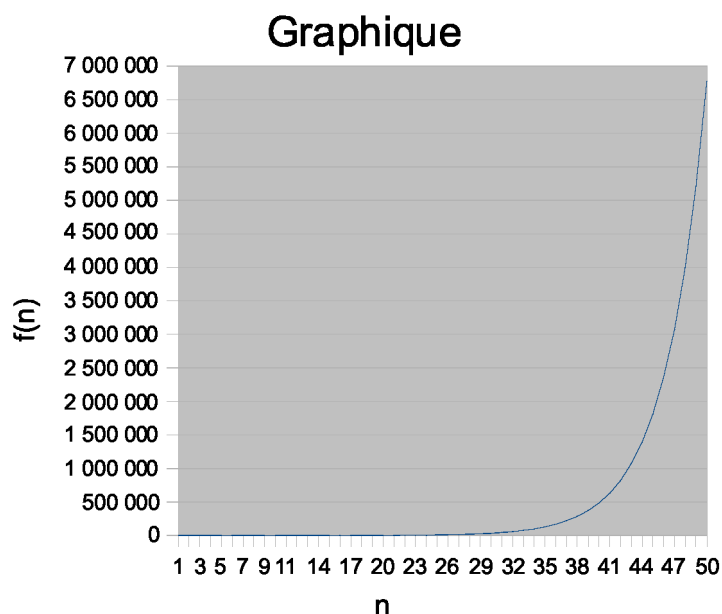
Cette méthode peut sembler longue et pénible mais tout ceci est effectué par l'ordinateur. On peut ainsi former de très grands mots.

Nombre de mots d'une longueur donnée

Désormais, on notera $f(n)$ le nombre de mots corrects de longueur n . Par exemple :

$$f(2) = 6$$

$$f(60) = 94\,766\,136 \quad \textcolor{yellow}{(3)}$$



Propriétés de $f(n)$

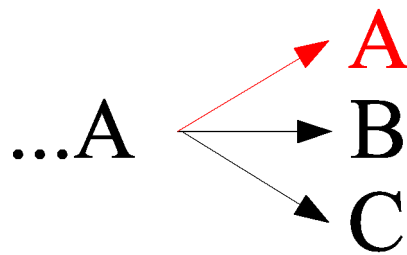
1) $f(n)$ est divisible par 6

Car à partir d'un mot correct on peut en former 5 autres par permutations. En remplaçant chaque syllabe par une autre.

A	B	C
A	C	B
B	A	C
B	C	A
C	B	A
C	A	B

2) $f(n+1) \leq 2 \times f(n)$

En effet si un mot se termine par A donc s'écrit sous la forme xA on peut lui rajouter dans le meilleur des cas soit un B soit un C car la séquence «AA» est fausse.



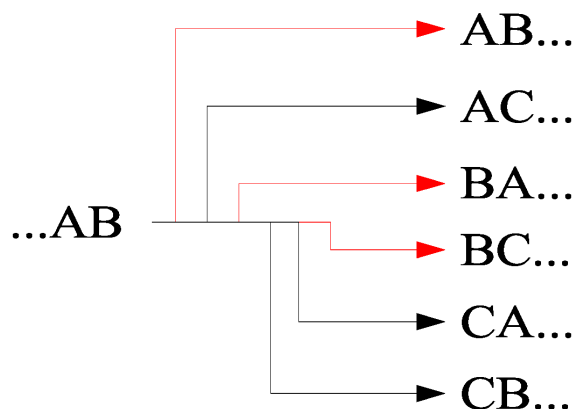
$$3) f(m+n) \leq \frac{f(m) \times f(n)}{2} \quad (\text{pour } n \text{ et } m \text{ supérieurs à } 2)$$

Démonstration

On peut former un mot de longueur $m+n$ en concaténant un mot de longueur m et un mot de longueur n .

Exemple : ABCAB + ACA → ABCABACA

Mais il y a des répétitions obligatoires dans au moins la moitié des cas. En effet il découle de la propriété 2 (4) qu'il y a autant de mots qui commencent par AB que de mots qui commencent par AC, BA, BC, CA ou CB. De plus il y a donc autant de mots qui se terminent par AB que par AC, BA, BC, CA ou CB. Donc si par exemple le mot de longueur m se termine par AB, le mot de longueur n peut commencer par AB, AC, BA, BC, CA, CB. Donc au moins la moitié des mots formés ainsi sont interdits.



De même pour les mots qui commencent par AC, BA, BC, CA ou CB. On a donc bien :

$$f(m+n) \leq \frac{f(m) \times f(n)}{2}$$

Réponses aux questions

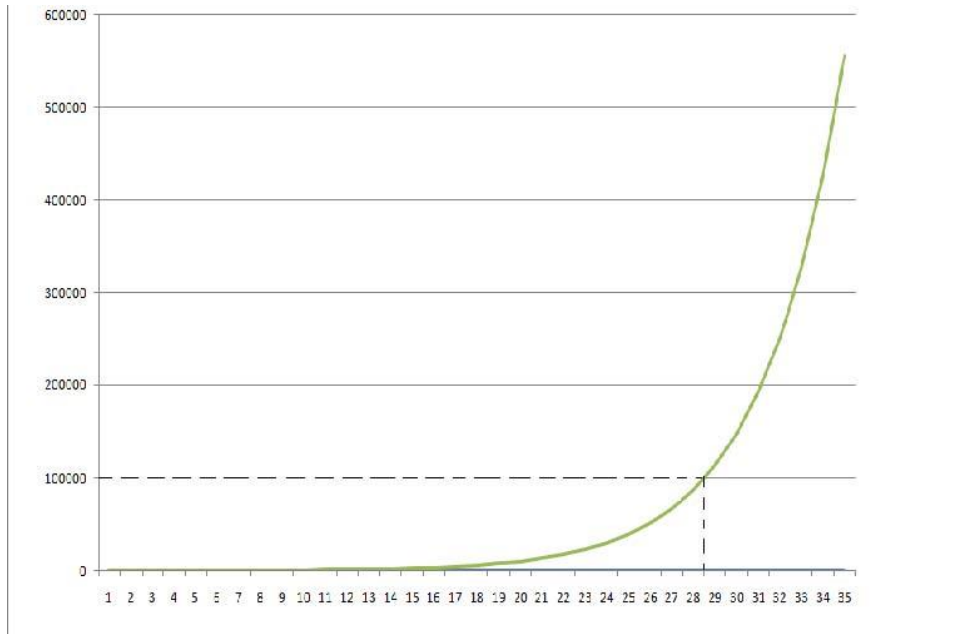
Existe-t-il des mots de 100, 1 000 ou 100 000 syllabes ?

Oui un programme informatique nous a permis de trouver des mots de 100 000 syllabes et bien plus encore.

Quel est le nombre minimum de syllabes pour générer une langue de 100 000 mots comme le français ?

Le nombre de syllabes du plus long mot serait alors de 29. (5)

Nombre de mots cumulés (6)



Résultats intéressants

Transformations de 2 syllabes

Nous avons essayé de trouver des transformations valides de 2 syllabes, c'est à dire qu'on associe à une syllabe d'un mot correct deux syllabes pour former un autre mot correct deux fois plus long. Nous avons montré qu'il n'existe pas de transformation valide de 2 lettres. En effet si par exemple :

A → **AB**

B → **AC**

C → **BC**

On a:

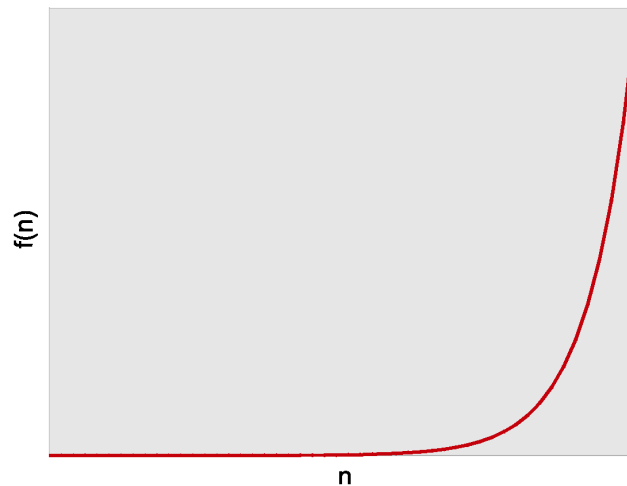
AC → **AB BC**

Ce qui est **interdit**. On peut procéder de même pour les transformations de plus de lettres.

Approximation de $f(n)$

On remarque que f ressemble à une fonction exponentielle. On essaye donc d'approximer $f(n)$ par $b \times a^n$.

Graphique



Si $f(n) = b \times a^n$ alors :

$$f(n+m) = b \times a^{n+m} = b \times a^n \times a^m = \frac{f(m) \times f(n)}{b}$$

Donc $b \geq 2$ car on a démontré que : $f(m+n) \leq \frac{f(m) \times f(n)}{2}$

De même on a : $f(n+1) = b \times a^{n+1} = b \times a^n \times a$

Et : $2 \times f(n) = 2 \times b \times a^n$

Or comme $f(n+1) \leq 2 \times f(n)$,

on a : $b \times a^n \times a \leq 2 \times b \times a^n$

Et donc : $a \leq 2$

On obtient donc a en faisant :

$$\frac{f(n)}{f(n-1)} \approx \frac{ba^n}{ba^{n-1}} = \frac{a^n}{a^{n-1}} = a^{n-n+1} = a$$

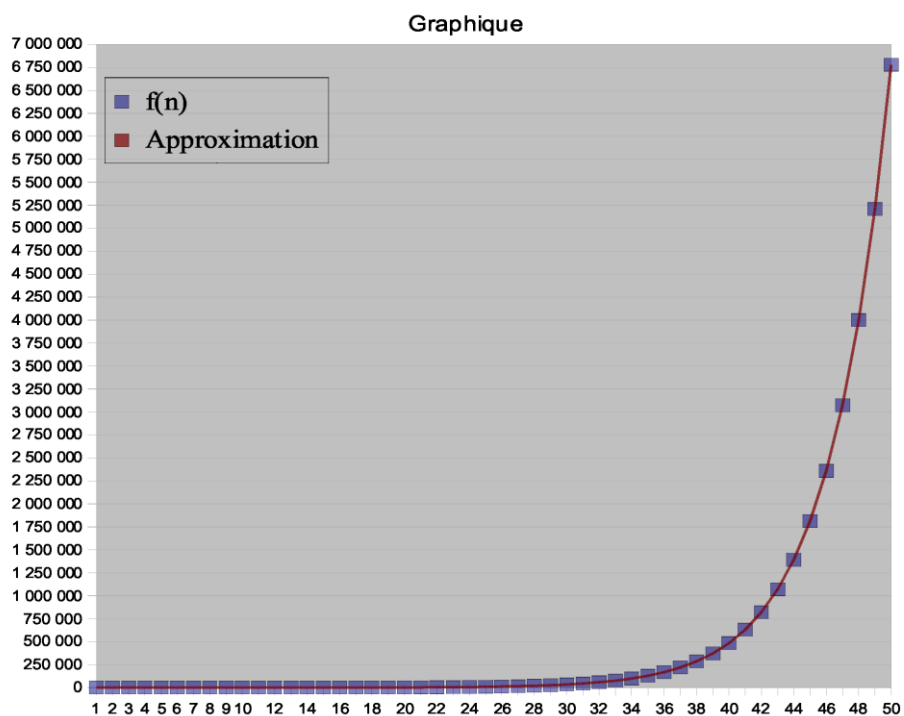
De même on obtient b en faisant :

$$\frac{f(n)}{a^n} \approx \frac{ba^n}{a^n} = b$$

On observe que $f(n) / f(n-1) = a$ converge vers une valeur proche de 1,3018.

Voici un tableau de l'approximation de $f(n)$ pour $n \leq 50$.

f_n	$12,7 * 1,3018^n$	$f(n)/f(n-1) = a$	b
1	3	17	/
2	6	22	2,000000000000000 3,54
3	12	28	2,000000000000000 5,44
4	18	36	1,500000000000000 6,27
5	30	47	1,666666666666667 8,02
6	42	62	1,400000000000000 8,63
7	60	80	1,42857142857143 9,47
8	78	105	1,300000000000000 9,46
9	108	136	1,38461538461538 10,06
10	144	178	1,33333333333333 10,3
11	204	231	1,416666666666667 11,21
12	264	301	1,29411764705882 11,14
13	342	392	1,29545454545455 11,09
14	456	510	1,33333333333333 11,36
15	618	664	1,35526315789474 11,83
16	798	864	1,29126213592233 11,73
17	1 044	1 125	1,30827067669173 11,79
18	1 392	1 464	1,33333333333333 12,07
19	1 830	1 906	1,31465517241379 12,19
20	2 388	2 481	1,30491803278689 12,22
21	3 180	3 230	1,33165829145729 12,5
22	4 196	4 205	1,31949685534591 12,67
23	5 418	5 474	1,29122974261201 12,57
24	7 032	7 126	1,29789590254707 12,53
25	9 198	9 277	1,30802047781570 12,59
26	11 892	12 077	1,29288975864318 12,51
27	15 486	15 722	1,30221997981837 12,51
28	20 220	20 466	1,30569546687331 12,55
29	26 424	26 643	1,30682492581602 12,6
30	34 422	34 684	1,30267938237965 12,6
31	44 862	45 152	1,30329440474115 12,62
32	58 446	58 779	1,30279523873211 12,63
33	76 122	76 518	1,30243301509085 12,63
34	99 276	99 611	1,30416962244818 12,66
35	129 516	129 674	1,30460534268101 12,68
36	168 546	168 810	1,30135272862040 12,68
37	219 516	219 756	1,30241002456303 12,69
38	285 750	286 079	1,30172743672443 12,69
39	372 204	372 418	1,30255118110236 12,69
40	484 446	484 813	1,30156043460038 12,69
41	630 666	631 130	1,30182930605269 12,69
42	821 154	821 605	1,30204260258203 12,69
43	1 069 512	1 069 565	1,30244996675410 12,7
44	1 392 270	1 392 360	1,30178062518233 12,7
45	1 812 876	1 812 574	1,30210088560409 12,7
46	2 359 710	2 359 609	1,30163894276277 12,7
47	3 072 486	3 071 739	1,30206084645995 12,7
48	4 000 0023	998 790	1,30187802320336 12,7
49	5 207 7065	205 624	1,30192584903708 12,71
50	6 778 926	6 776 682	1,30171058043599 12,7



Approximation avec $a=1,3018$ et $b=12,7$ (7)

Les mots finis (8)

On peut démontrer que tous les mots finis sont du type : $zyxABxACyxABxA$

Avec x,y,z des groupes de syllabes corrects pouvant être nuls. (9)

Notes d'édition

- (1) Les trois itérations correspondent aux trois syllabes considérées (A, B et C).
- (2) Cet algorithme peut également être adapté pour générer tous les mots d'une longueur n fixée.
- (3) La valeur de $f(2)$ peut facilement être vérifiée à la main en énumérant tous les mots corrects. En revanche, la valeur de $f(60)$ semble avoir été calculée à l'aide d'un programme informatique.
- (4) C'est en fait la Propriété 1 (concernant les permutations) qui est utilisée.
- (5) La question considérée ici est probablement plus précisément la suivante : "Si les Shadocks ont une langue de 100 000 mots (comme le français), que peut-on dire de la longueur du plus grand mot ? »
Et la réponse est que le plus long mot a au moins 29 syllabes.
- (6) La Figure « Nombre de mots cumulés » représente la longueur du plus long mot généré (axe des abscisses) en fonction du nombre total de mots générés (axe des ordonnées).
- (7) La Figure « Approximation avec $a = 1,3018$ et $b = 12,7$ » représente le nombre de mots de longueur n avec n compris entre 0 et 50, et la fonction continue : $x \mapsto 12,7(1,3018)^x$.
- (8) Rappelons la définition des « mots finis » donnée en p 2 : ce sont les mots qui ne peuvent pas être prolongés.
- (9) On peut facilement vérifier que ces mots sont bien finis.
Attention toutefois cette proposition ne les décrit pas tous : par exemple elle ne décrit que des mots finis qui se terminent par la syllabe A. Il faut en fait rajouter tous ceux qu'on obtient en échangeant A, B et C.