CRYPTOGRAPHIE

Année 2007 - 2008

Amandine MILLET, élève de 6ème
Arthur THOMAS, Clémence GALLIOT, Abigaïl ROUSSEAU, Ella VETTER, Alice BELISSA, Thomas BRUDER, Vincent HUBERT, Guillaume BASILE, Aurélien BUSSON, Eléonor LINDER, Jérémy MOREAU, Léa STUDER, élèves de 4ème

Etablissement : Collège G CHEPFER, Villers les Nancy

Enseignantes : Louisette HIRIART, Christelle KUNC

Chercheur : Thomas CHAMBRION, université Henri Poincaré, Nancy

La cryptographie est la science qui étudie les moyens de rendre un message inintelligible à qui ne possède pas une certaine clé de lecture. Ses applications sont innombrables pour garantir la confidentialité des échanges, par exemple sur internet.

1 – Ce que Vercingétorix aurait aimé savoir faire : le code de César.

Le code de César est la méthode de cryptographie la plus ancienne communément admise par l'histoire. César, pour faire parvenir des messages à ses troupes, sans risquer qu'ils tombent aux mains des ennemis, décalait toutes les lettres de l'alphabet par exemple de 3 rangs.

Texte clair	A	B	C	D	E	F	G	H	I	J	K	J	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Texte codé	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*W*	*X*	*Y*	*Z*	*A*	*B*	*C*

Il n'y a que 25 façons différentes de crypter un message selon la méthode de César, puisqu'il y a 25 façons de décaler les lettres de l'alphabet qui contient 26 lettres. Cela en fait donc un code très peu sûr, puisqu'il est très facile de tester toutes les possibilités.

Notre chercheur nous avait donné un texte à décrypter selon cette méthode. Voici le début du texte :

WPACP XTPCB FTGTE FYNSL XPLFD PYQFT ENPEZ MUPEY ZFGPL FWPDP NZYOL AACZN SLWPE CZTDT XPZDL QLTCP FYWTN ZFAZF CWPOC ZXLOL TCPWL NNZFE FXLYN PLTYD CPYOE ZFEQL XTWTP CNPBF TYZFD ALCLT .... (il y avait en tout 541 lettres) ....

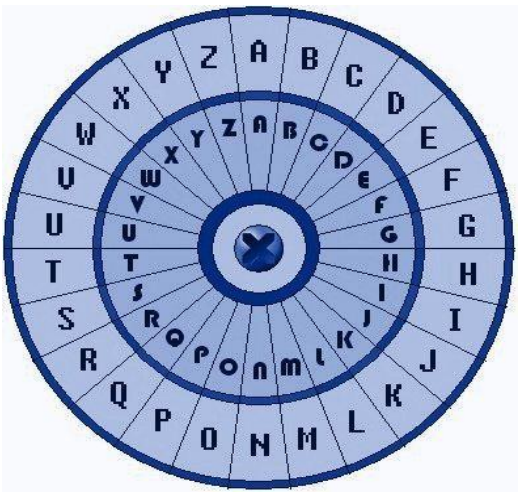
Nous nous y sommes tous mis, chaque élève du groupe prenant un décalage différent.

Cela a été assez rapide et dans la séance nous avons trouvé qu'il s'agissait d'une **fable de La Fontaine : «Le premier qui vit un chameau.... »**.

Le décalage de l'alphabet était de 15 rangs.

Texte codé	*W*	*P*	*A*	*C*	*P*	*X*	*T*	*P*	*C*	*B*	*F*	*T*	*G*	*T*	*E*	*F*	*Y*	*N*	*S*	*L*	*X*	*P*	*L*	*F*	*D*	*P*
Texte clair	**L**	**E**	**P**	**R**	**E**	**M**	**I**	**E**	**R**	**Q**	**U**	**I**	**V**	**I**	**T**	**U**	**N**	**C**	**H**	**A**	**M**	**E**	**A**	**U**	**S**	**E**

Ensuite nous nous sommes amusés à coder de petits messages et avons proposé au public de les décoder sur notre stand au congrès. Nous avons aussi construit des disques qui permettaient de rapidement trouver les correspondances de lettres selon le décalage des lettres de l'alphabet choisi.



Ensuite le groupe s'est scindé en deux, chacun essayant de décrypter l'un des deux autres textes donnés par notre chercheur selon deux méthodes différentes.

2 – Le chiffrement mono alphabétique

Dans un tel chiffrement, chaque lettre est substituée à une autre lettre de l'alphabet, chaque lettre du texte chiffré représentant toujours la même lettre du texte en clair.
Cela se complique sacrément, il y a 26 ! possibilités (1) (pour le A, il y a 26 choix possibles ; pour le B, il reste 25 choix possibles (2) ; pour le C, 24 choix ; etc....., il y a donc 26 x 25 x 24 x 23 x.....x 2 x 1 possibilités).
Cela représente environ 4×10^{26} possibilités (3), c'est énorme !

Le texte à décrypter était le suivant :

JMQID RROBQ MIORM UXSDO UXROD BJOUQ EUJDO OJTQO VKRHM DJRVK URMRR UQOMU EKUQS CUDGO QORTO SORQK HMDJR VKROJ JOHDR SOICU RSOGO UQVMD JOORL OQMJI ORKJT MGGOR ICOZL MGGMR LGOUQ OQGOU QDHUL DRRMJ IOHMD RPUOV KDREO VKURH OHODJ OUDOT OTKJJ OLGUR PUOAQ DTMJJ DIURL MQMDR ROZIK JRTOQ JOPUO LQORM BOMHO RYOUX IOTTO TQDRT ORROK ARIUQ OOTIO RRKHA QORQO BMQSR OQQMJ TMGMV OJTUQ OTKUT VKURQ DTGMN KQTUJ OKAOD TMVKR VKOUX JOQKJ JMQID RROIO JORTN MDTJO QKJOR TMHKU QOUXJ MQIDR ROVKU RJOQK JSOLU DRUJH KHOJT HMDRL KUQTK UTOHM VDOEM DHOPU OSDRE OMDHO QEDSK GMTQO EUJDO

Tout d'abord, nous avons cherché les fréquences d'apparition de chaque lettre dans le texte crypté pour les comparer à celles données dans le dictionnaire français.

Pour calculer par exemple la fréquence d'apparition du O dans le texte crypté, nous avons compté tous les O, il y en a 79 sur les 460 lettres du texte.

79/460 = 0,1717

La fréquence d'apparition du O est donc de 17,17%.

Fréquence d'apparition des lettres dans notre texte codé dans l'ordre décroissant :

O	*R*	*Q*	*M*	*J*	*U*	*D*	*K*	*T*	*H*	*I*	*G*	*V*	....
17,2%	10,8%	7,8%	7,8%	7,4%	7,1%	6,9%	5,8%	5,6%	3,5%	3%	2,6%	2,6%	.....

En général, les lettres françaises (comptées sans accent) apparaissent avec les fréquences suivantes :

E	A	S	I	T	N	R	U	L	O	D	C	P
17,1%	8,1%	7,9%	7,5%	7,2%	7,1%	6,6%	6,3%	5,5%	5,4%	3,7%	3,2%	3%

M	V	Q	F	B	G	H	J	X	Y	Z	W
3%	1,6%	1,4%	1,1%	0,9%	0,9%	0,7%	0,5%	0,4%	0,3%	0,1%	0,1%

En comparant les tableaux, nous avons déduit que **le *O* du texte codé correspondait au E** dans le texte décrypté, ces deux lettres ayant pratiquement la même fréquence d'apparition dans les deux textes. Pour les autres lettres, nous ne pouvions rien conclure.

Il nous fallait trouver d'autres éléments. Nous avons alors repris le texte codé et cherché les lettres doubles en supposant que cela devait être des consonnes doubles comme nous en avons en français (tt, ss, nn, rr et ll) et que dans ce cas, elles étaient entourées de voyelles.

Nous avons relevé dans le texte codé :

KJ J O-OJ JO -MJ J D MG G O -MG G M OT TO

OQ Q M OR R K -OR R O -MR R U -DR R M -DR R O

Nous avons alors déduit que les lettres ***O***, ***M***, ***U***, ***D*** *et* ***K*** du texte codé devaient être les 5 voyelles, et que les lettres ***R***, ***Q***, ***J***, ***T*** *et* ***G*** du texte codé devaient être des consonnes les plus fréquentes et souvent doublée en français. On a alors repris les 2 tableaux de fréquence et fait des rapprochements.

Fréquences d'apparition des lettres dans un texte en français

E	A	S	I	T	N	R	U	L	O	D	C	P	.....
17,1%	8,1%	7,9%	7,5%	7,2%	7,1%	6,6%	6,3%	5,5%	5,4%	3,7%	3,2%	3%	.....

O	*R*	*Q*	*M*	*J*	*U*	*D*	*K*	*T*	*H*	*I*	*G*	*V*	....
17,2%	10,8%	7,8%	7,8%	7,4%	7,1%	6,9%	5,8%	5,6%	3,5%	3%	2,6%	2,6%	.....

Fréquences d'apparition des lettres dans notre texte codé

Les flèches indiquent les correspondances dont nous étions pratiquement sûrs.

Nous concluons aussi que :

le ***U*** et le ***D*** du texte codé correspondent aux voyelles **I** ou **U.**

le ***Q***, le ***J*** et le ***T*** du texte codé correspondent aux consonnes **T, N** ou **R.**

Nous n'avons pas trouvé d'autres associations de lettres, et ensuite nous avons essayé les combinaisons possibles jusqu'à ce que des mots apparaissent et nous permettent de faire les associations de lettres.

Ainsi, le ***U*** du texte codé correspond au **U** et le **D** du texte codé au **I.**
Et les consonnes ***Q***, ***J*** et ***T*** du texte codé correspondent respectivement aux consonnes **R**, **N**, et **T.**

C'est en essayant de nombreuses combinaisons et en essayant de trouver des mots courants que nous avons fini par décrypter ce texte, mais cela a été très long et laborieux. Nous avons trouvé seulement à la mi-mars. Il s'agissait de la tragédie **Britannicus, acte 2 scène 2 de Racine.**

Remarques :
Le décryptage aurait été facilité si le découpage des mots avait été respecté, car en français :
- le A est la lettre isolée la plus fréquente
- les mots de 2 lettres EN, NE, SE et TE apparaissent souvent.

Il faut un texte de longueur raisonnable à décrypter afin que les statistiques soient fiables.

3 – Le chiffre de Vigenère

a) Tout d'abord, il a fallu comprendre cette méthode de chiffrement. Coder selon Vigenère : on choisit un mot clé, par exemple **CHAT.**
Pour crypter le texte :
- on décale la première lettre du texte de 2 rangs (C est le 3ème lettre de l'alphabet et correspond à un décalage de 2 rangs)
- on décale la deuxième lettre de 7 rangs (H correspond à un décalage de 7 rangs)
- la troisième lettre reste identique car A est la première lettre de l'alphabet)
- on décale la quatrième lettre de 19 rangs (T correspond à un décalage de 19 rangs)
- quand on a fini la clé, on recommence au début, on décale la cinquième lettre du texte à chiffrer de 2 rangs, la sixième de 7 rangs, etc...

Exemple :

Texte initial	**M**	**E**	**S**	**S**	**A**	**G**	**E**	**C**	**O**	**D**	**E**
Rang de décalage	2	7	0	19	2	7	0	19	2	7	0
Texte crypté	***O***	***L***	***S***	***L***	***C***	***N***	***E***	***M***	***Q***	***K***	***E***

On constate alors que la même lettre **S** du texte initial est codée par deux lettres différentes et aussi que la même lettre ***L*** du texte crypté correspond à deux lettres différentes du texte initial.

b) Texte à décrypter donné par notre chercheur

1ère étape : Chercher la longueur de la clé. Pour trouver la longueur de la clé, on a repéré des séquences de trois lettres (ou plus) répétées dans le texte codé et on s'est dit que ces répétitions ne devaient pas être fortuites. Si une séquence de trois lettres est répétée dans le message codé avec une distance « d », on peut se dire qu'il s'agit peut-être de la même séquence de trois lettres du texte initial, codée avec la même séquence de lettres de la clé. Par conséquent, si « c » est la longueur de la clé, pour que les séquences soient codées avec les mêmes lettres de la clé, il faut que « c » divise « d ». Et par suite, « c » est égal au plus grand diviseur commun des distances des séquences répétées du texte codé. (4)

Dans notre texte codé nous avons repéré les séquences répétées de couleurs différentes, puis nous avons compté la distance entre chaque séquence répétée.

Notre texte codé était :

EVVQG JYPLV MVERW JHJYF NDIPW VFCFZ YWQGU TMARAG XTUTV
YUENE JGGJY GGVNT XJCDB VHRHL LXNHO GVVOZ FVGGJ
LCUHE NNXJJ NNJXK YWCEB CYUTT IPMVH VXIYP MFOVX ROVKV WJHJY
PHENR HZHVV FOVND YFXEX GLZLG KGFWL HOKEJ YPHEN GGHOQ BZFPX
JNRTJ PTTZM GFSFC UCYSN VNQNJ MGMII OIVHV FRCUI COVHK WGERN
FFCI GVKWX CURNZ MUTEW GWWVVK
XEDWZ VLGMU CUMZH INVLN XMLCB UUXXT FGYRO ZJLCG LKJTH GLGFV
HVVVK WHEHQ FDYNX SIPLV HUHLF CKRCU HEYUM EUVNI YNEVG GGKYI
TCYGG KIWLC YUAFG OXJYV TZHUB HOGER XKOVL UBKYF XEIUH GCPBF
HUGVP KXENR TJXGV VKWXC YUNEM UHENR ELMTT ZMQGE UDEVM SNVFG
LROVK VMOTZ MUXLF GFVHV WVWGJ LYPHL MEHEX WBJIP LEIUI VHUXJ
JCKUC XXIMG LMIKX JYVGV WQGJC FXIIP LGUUE VMOXD YUVYI UXJWC
KTYPX JNRTJ UULVT FTMIK KCYUI ICVUF HOTZM NXGLK GTCRT CYUMU
YNTGJ NBHOG
KSCGG

Séquences répétées	Distance « d » entre les deux séquences
JHJ	28 = 4 x 7
HEN	48 = 4 x 12
PHEN	36 = 4 x 9
VHV	92 = 4 x 23
GGK	8 = 4 x 2
HOG	232 = 4 x 58
ENR	20 = 4 x 5

Nous avons donc supposé que **la clé était un mot de quatre lettres** puisque toutes les distances entre les séquences répétées sont des multiples de 4.

2ème étape : Le découpage du texte en quatre séries de lettres.

Nous avons reconstruit alors quatre séries de lettres :

- l'une avec les 1ère, 5ème, 9ème, etc... lettres du texte codé : ***EGLE.....***
- une autre avec les 2ème, 6ème, 10ème, etc.... lettres du texte codé : ***VJVR.....***
- une troisième avec les 3ème, 7ème, 11ème, etc ... lettres de texte codé : ***VYMW.....***
- enfin une dernière avec les 4ème, 8ème, 12ème, etc... lettres du texte codé : ***QPVJ.....***

Dans chaque série de lettres, il y a le même décalage de l'alphabet, pour le trouver, il suffit de repérer la lettre la plus fréquente qui correspond à un **E** dans la langue française. Lorsque la lettre la plus fréquente n'apparaît pas immédiatement, on effectue une analyse fréquentielle rapide. On connait alors le décalage des lettres de l'alphabet et on peut décrypter les séries de lettres selon la méthode de César.

3ème étape : Etude de chaque série de lettres.

- Dans la 1ère série : ***E G L E.....*** La lettre la plus fréquente était le ***X*** qui représente le **E**, ce qui nous a permis de trouver le décalage de 7 rangs entre les lettres de la série et les lettres correspondantes décryptées :

Texte codé	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Texte décrypté	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z

Ainsi *E G L E* .... correspondent à L N S L.....
On fait de même pour chacune des trois autres séries de lettres.

- Dans la 2ème série : ***V J V R .....***
La lettre la plus fréquente est ***V*** qui représente la lettre **E,** d'où le décalage de 9 rangs entre les lettres de la série et les lettres correspondantes décryptées.

Texte codé	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Texte décrypté	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z

Ainsi *V J V R* ... correspondent à E S E A ...

- Dans la 3ème série : ***V Y M W .....***
La lettre la plus fréquente est ***Y*** qui représente la lettre **E,** d'où un décalage de 6 rangs entre les lettres de la série et les lettres correspondantes décryptées.

Texte codé	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Texte décrypté	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z

Ainsi *V Y M W* ... correspondent à B E S C ...

- Dans le 4ème série: ***Q P V J.....***
La lettre la plus fréquente est ***G*** qui représente la lettre **E**, d'où un décalage de rangs entre les lettres de la série et les lettres correspondantes décryptés.

Texte codé	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
Texte décrypté	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z

Ainsi Q P V J ... correspondent à O N T H...

Il ne nous reste plus qu'à rassembler les lettres décryptées de chaque série
L N S L ...
E S E A ...
B E S C ...
O N T H ... dans l'ordre de départ.
On trouve alors : **« LE BON SENS EST LA CHOSE ... »**
Ensuite « google » a cherché pour nous la suite du texte crypté qui était le début du **DISCOURS DE LA METHODE de Descartes.**

Dernière étape : Retrouver la clé (qui ne sert plus à rien)
Nous avons cherché à quelles lettres correspondaient les **A** du texte décrypté dans chacune des trois séries de lettres et en respectant l'ordre des séries, nous avons trouvé la clé : **T R U C.**

Remarques

Pour le décryptage il est nécessaire que le texte soit suffisamment long pour repérer des séquences de 3 lettres répétées et pour avoir une analyse fréquentielle correcte dans chacune des séries de lettres.

Nous avons passé beaucoup de temps à faire une grille pour décrypter tout le texte en entier alors qu'en fait, comme il s'agissait d'un texte d'un auteur classique, seuls les six premiers mots nous ont été indispensables, ces premiers mots « le bon sens est la chose » ont suffi à identifier le texte grâce à la recherche informatique. Nous avons tout de même vérifié la suite. Pour un message codé inventé par un camarade, c'est beaucoup plus difficile.

Notes d'édition

(1) 26 ! se lit « factorielle de 26 » et désigne le produit des nombres entiers de 1 à 26 :
1 x 2 x 3 ... x 25 x 26.

(2) « *il reste 25 choix possibles* » : la lettre par laquelle on remplace A ayant déjà été fixée.

(3) 4×10^{26} : signifie que ce nombre s'écrit avec 27 chiffres, le premier étant un 4.

(4) En tout cas, il doit s'agir d'un diviseur commun, pas forcément du plus grand. En pratique, il est effectivement vraisemblable que ce soit le plus grand dans un grand texte où les occurrences de la séquence de trois lettres sont nombreuses.