MATHÉMATIQUES ET SCIENCES HUMAINES

G. TH. GUILBAUD

Note sur les comptabilités markoviennes

Mathématiques et sciences humaines, tome 66 (1979), p. 99-112

<http://www.numdam.org/item?id=MSH_1979__66__99_0>

Math. Sci. hum. (17^{e} année, n°66, *1979*, p.99-112)

NOTE SUR LES COMPTABILITES MARKOVIENNES

G.Th. GUILBAUD (1)

0. Une succession de signes (ou : chaîne, ou : séquence, ou : suite, ou encore : chronique, comme il vous plaira). Mettons, pour commencer, qu'il n'y a que deux signes disponibles. Et appelons-les : C et V,"Consonne" et "Voyelle" à seule fin d'évoquer les premières expériences faites naguère par A.A Markov.

Soit, pour exemple, la transcription (ou : codage) du texte qui précède (en négligeant intervalles et ponctuations) :

VCVCVCCVCCVVCCVCVCCVCVVCCVVCVVVCVCVVCCVV ....

arrêtons-nous ici (quarante signes) et comptons.

Il y a 21 fois V et 19 fois C.

1. Mais on doit aussi compter les sous-séquences. Et d'abord :

VC	apparaît	13 fois
CV	"	13 "
CC	"	6 "
VV	"	7 "
	Total :	39 doublets (pour 40 signes).

(1) Centre de Mathématique Sociale de l'Ecole des Hautes Etudes en Sciences Sociales.

On remarque que les CV et les VC sont également nombreux. On peut trouver que c'est évident. Mais en prévision de ce qui va suivre, on fera bien de le démontrer.

Si VC apparaît 13 fois et VV 7 fois, cela signifie que V apparaît (13 + 7) = 20 fois *suivi* de quelque signe.
Avec des notations faciles à comprendre, on peut écrire :

$$(VC) + (VV) = (V*)$$

et de même

$$(CV) + (VV) = (*V)$$

(le signe * désigne n'importe lequel des deux signes : V ou C).

Il en résulte :

$$(VC) - (CV) = (V*) - (*V)$$

Ici, pour notre exemple, le nombre (V*) des V *suivis* de quelque chose est égal au nombre (*V) des V *précédés* de quelque chose, parce que la séquence étudiée commence et finit par le même signe.

Précisons encore, pour préparer ce qui va suivre.
Ecrivons :

$$(V) = (V*) + f(V)$$

avec : f(V) = 1, si la séquence finit par V, et f(V) = 0, sinon ;
ou bien encore : f(V) = nombre.de V en *fin* de séquence
et : d(V) = nombre de V (soit zéro, soit un) en *début*.

$$(V) = (*V) + d(V)$$

D'où :

$$(V*) - (*V) = (V) - f(V) - (V) + d(V)$$
$$(V*) - (*V) = d(V) - f(V)$$

Et enfin :

$$(VC) - (CV) = d(V) - f(V)$$

Le second membre étant nul pour les séquences :

V V
comme pour C C

Pour les séquences : C .....V, on a d(V) = 0 et f(V) = 1
et pour : V .....C, on a d(V) = 1 et f(V) = 0.

Ainsi est établie la première proposition :
La différence (VC) - (CV) entre le nombre des CV et des VC, est nécessairement égale soit à 0, soit à +1, soit à -1.

Donc si quelqu'un me dit qu'il a compté 125 fois CV et 123 fois VC, je suis sûr qu'il a commis au moins une erreur.
On peut même aller plus loin. Si une comptabilité se présente comme correcte, par exemple :

CV = 7 fois VC = 8 fois
CC = 2 fois VV = 4 fois

on peut chercher à reconstituer la séquence.

Placer d'abord les CV et VC :

VCVCVCVCVCVCVCVC

Soit : $(VC)^8 = V(CV)^7$

ensuite, choisissons deux signes C quelconques pour les doubler, et quatre signes V qu'on remplacera par VV.

Cela peut se faire de nombreuses façons :

pour C de $\frac{8 \times 9}{2} = 36$ façons

pour V de $\frac{8 \times 9 \times 10 \times 11}{1 \times 2 \times 3 \times 4} = 330$ façons

au total $36 \times 330 = 11880$ façons.

Finalement : PROPOSITION 1.
La condition nécessaire et suffisante pour qu'une statistique
(CV) = ... (VC) = ... (CC) = ... (VV) = ...
soit réalisable, est que la différence (CV) - (VC) soit égale à zéro ou bien à plus ou moins l'unité.

2. Passons maintenant aux sous-séquences de trois signes (triplets).
En reprenant notre exemple du début, on trouve ceci : (ceux des lecteurs qui sont assidus essaieront de vérifier cette comptabilité, ne serait-ce que pour mesurer la peine qu'on a quand l'on veut éviter les erreurs)

VCV 7 fois
CVC 7 fois
VCC 6 fois
CCV 6 fois
VVC 5 fois
CVV 6 fois
CCC 0 fois
VVV 1 fois

On a envie de vérifier ; appliquons la même méthode que précédemment et les mêmes notations :

(VC*) = (VCV) + (VCC) = 7 + 6 = 13
(*VC) = (VVC) + (CVC) = 5 + 7 = 12
(VC) = (VC*) + f(VC) = 13 + 0 = 13
(VC) = (*VC) + d(VC) = 12 + 1 = 13

f signifie : "fin"
et d : "début"

Cette première vérification est bonne.

Il y a d'autres vérifications :

d'abord : (CV*) = (CVV) + (CVC) = 6 + 7 = 13
(*CV) = (VCV) + (CCV) = 7 + 6 = 13
(CV) = (*CV) + d(CV) = 13 + 0 = 13
(CV) = (CV*) + f(CV) = 13 + 0 = 13

puis : (CC) = (*CC) + d(CC) = (CCC) + (VCC) + d(CC) = 0 + 6 + 0 = 6
(CC) = (CC*) + f(CC) = (CCC) + (CCV) + f(CC) = 0 + 6 + 0 = 6

enfin : (VV) = (*VV) + d(VV) = (VVV) + (CVV) + d(VV) = 1 + 6 + 0 = 7
(VV) = (VV*) + f(VV) = (VVV) + (VVC) + f(VV) = 1 + 5 + 1 = 7

Tout a bien marché.
Reste à énoncer, d'une façon générale, les conditions nécessaires. Et aussi à montrer qu'elles sont suffisantes.

3. En reprenant les diverses équations écrites ci-dessus :

(VC) = (VCV) + (VCC) + f(VC)
(VC) = (VVC) + (CVC) + d(VC)
(CV) = (CVV) + (CVC) + f(CV)
(CV) = (VCV) + (CCV) + d(CV)

d'où : $(VCV) + (VCC) + f(VC) = (VVC) + (CVC) + d(VC)$

$(CVV) + (CVC) + f(CV) = (VCV) + (CCV) + d(CV)$

ou bien : $(VCV) - (CVC) = (VVC) - (VCC) + d(VC) - f(VC)$

$(VCV) - (CVC) = (CVV) - (CCV) + f(CV) - d(CV)$

ce qui signifie que les différences :

$(VCV) - (CVC)$, $(CVV) - (CCV)$, $(VVC) - (VCC)$

sont toujours "presque" égales (à une unité près, puisque les nombres (f-d) sont égaux soit à 0, soit à +1, soit à -1).

Mais on a déjà vu que (VC) et (CV) sont toujours aussi "presque" égaux. Il en résultera la presque égalité de (CVV) et (VVC) d'une part, de (VCC) et (CCV) d'autre part.

On obtient donc d'abord ceci (en laissant tomber les f et d , provisoirement) :

$(CVV) \simeq (VVC)$

$(VCC) \simeq (CCV)$

$(VCV) - (CVC) \simeq (CVV) - (CCV) \simeq (VVC) - (VCC)$

$\simeq$ signifiant soit l'égalité exacte, soit la presque égalité à une unité près.

Vérifions d'abord ces relations sur quelques exemples empruntés à l'étude de Mme Petruszewycz (ci-dessus page 84) :

		H_p	P_p	H_v	P_v
1)	CVC	4831	5089	4861	4940
2)	VCV	3527	3927	3294	3175
3)	VCC	2180	1949	2333	2474
	CCV	2180	1950	2333	2474
4)	VVC	876	787	765	709
	CVV	876	787	765	709
différence entre 1) et 2)		1304	1162	1567	1765
différence entre 3) et 4)		1304	1162 ou 1163	1568	1765

4. Reste à montrer que ces conditions suffisent pour qu'on puisse réaliser effectivement une séquence.
On aura d'abord noté que les deux nombres (CCC) et (VVV) n'interviennent pas. C'était à prévoir : étant donné une séquence quelconque, on peut, à volonté, modifier les nombres de triplets CCC et VVV, sans toucher aux six autres nombres. Par exemple, les transformations :

$$\ldots\ VCCV\ \ldots\ \overset{\leftarrow}{\rightarrow}\ \ldots\ VCCCCV\ \ldots$$
$$\ldots\ VCVVCV\ \ldots\ \overset{\rightarrow}{\leftarrow}\ \ldots\ VCVVVVVCV\ \ldots$$

On peut donc se borner à traiter le cas où le nombre des CCC et des VVV est nul.

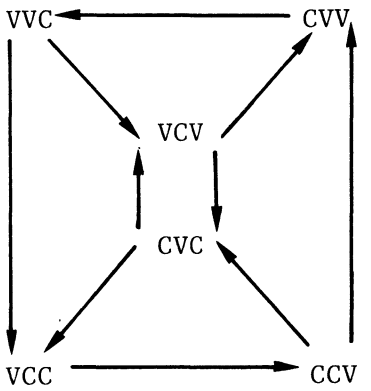
Et contentons-nous, pour éviter de trop longues écritures, d'un exemple numérique (qui s'inspire de la première colonne du tableau précédent).

Soit à montrer qu'on peut réaliser (de bien des façons) une séquence telle que :

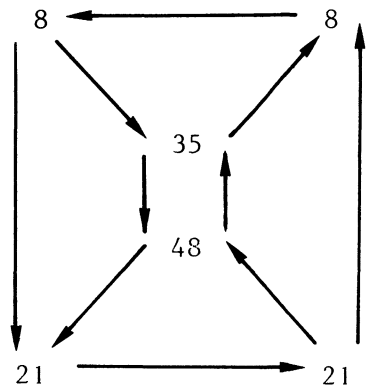
(CVC) = 48
(VCV) = 35
(VCC) = (CCV) = 21
(VVC) = (CVV) = 8

(qui satisfait bien à la condition : 48 - 35 = 21 - 8).

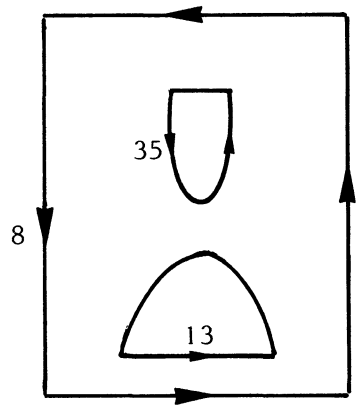
Il suffit de chercher l'enchaînement des divers triplets proposé par le programme. Mais cet enchaînement est soumis à des contraintes : ainsi un triplet CVC ne peut être "suivi" que de : VCC, ou de : VCV.
L'ensemble de ces contraintes est résumé par le diagramme ci-dessous.



Le problème est alors ramené à celui-ci : trouver un parcours qui passe 48 fois au noeud noté CVC, 35 fois à celui noté VCV, etc.



Une décomposition en cycles (possible grâce aux conditions) est évidente :



On va donc réaliser, d'abord séparément, les trois séquences suivantes :

1°) 8 fois : CVV, VVC, VCC et CCV

soit : (CVVCCVVCCVVCC ..... CVVCCV) = $(CVVC)^{8}CV$

2°) 13 fois : CVC, VCC et CCV

soit : (CVCCVCCVC ..... CVCCV) = $(CVC)^{13}CV$

3°) 35 fois : CVC et VCV

soit : (CVCVCVC ..... CVCV) = $(CVC)^{35}V$

Enfin on n'aura aucune peine à mettre les trois chaînes bout à bout : on a, en effet pris la précaution qu'elles commencent toutes (et finissent) de la même manière. Mais rien n'empêche de couper l'une en morceaux pour en intercaler une autre : on n'a que l'embarras du choix, et donc diverses façons de réaliser une chaîne unique conforme à la statistique donnée.

5. Pour traiter le petit exercice qu'on vient de voir, on s'est donné quelques facilités : les trois conditions, à savoir :

$$(VCC) = (CCV)$$
$$(VVC) = (CVV)$$
$$(CVC) - (VCV) = (VCC) - (VVC)$$

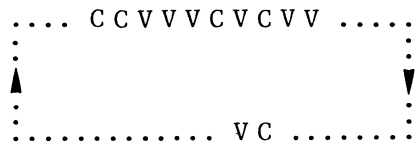
étaient réalisées exactement (et non pas à une unité près) comme on a vu que c'était possible.
Cette unité, en plus ou en moins, est un ennui permanent pour qui veut faire une théorie rigoureuse. Mais il y a une méthode pour écarter cette difficulté. Elle est classique en divers domaines (la statistique des chroniques en particulier), elle consiste à considérer des séquences *circulaires*.

Si l'on a, par exemple, la séquence [s]

[s] = [VCVCVV ... CV ... CCVV]

on lui associera la séquence circulaire [σ]

[σ] =



qui n'a plus ni fin (f) ni début (d), et pour laquelle toutes les conditions deviennent *exactes*.

Bien entendu, à une séquence circulaire [σ] correspondent plusieurs séquences linéaires [s] autant que de façons de pratiquer une coupure en [σ]. Mais la comptabilité pour les [s] se déduit facilement de la comptabilité pour [σ].

Ainsi, dans le cas figuré ci-dessus, la coupure qui transforme (σ) en (s) s'effectue ainsi : ... CVV | VCV ...
On voit alors qu'il y a *un* VVV et *un* VVC qui appartiennent à [σ] sans appartenir à [s].

On a donc : $(VVV)_s = (VVV)_\sigma - 1$
$(VVC)_s = (VVC)_\sigma - 1$

mais pour tous les autres triplets, l'égalité exactement.

D'une façon générale, un nombre de triplets sera le même pour [s] et [σ], sauf justement pour les *deux* triplets qui sont divisés par la coupure.

Désormais, pour alléger le discours, on ne parlera plus que des séquences circulaires.

Pour elles, on peut énoncer la PROPOSITION 2 :

Les conditions nécessaires et suffisantes pour pouvoir réaliser une séquence sont : (VCC) = (CCV)

(VVC) = (CVV)

(CVC) - (VCV) = (VCC) - (VVC)

On notera : trois relations pour huit variables.

6. Examinons maintenant le cas d'une statistique portant sur les quadruplets (suites de quatre signes).
Seize quadruplets peuvent se présenter. Empruntons encore à l'étude citée (ci-dessus page 85) une statistique concernant les quadruplets dans une séquence de 15000 signes (prose de Pouchkine) ; cet exemple a été choisi ici parce que le texte considéré finit comme il commence, et par conséquent, du point de vue statistique, fonctionne comme s'il était circulaire.

Dans le tableau ci-dessous on a rangé les quadruplets de signes dans l'ordre décroissant des fréquences, et on écrit aussi quelques unes des différences. Il saute aux yeux que plusieurs différences sont égales entre elles. Même à un oeil non averti, il apparaîtrait que ce ne sont pas là coïncidences fortuites. Ce sont en effet des équations démontrables, par la méthode déjà vue.
Par exemple : (VCVC) - (CVCV) = (VVCV) - (VCVV)

se déduit de : (VCVC) + (VCVV) = (VCV*)

(CVCV) + (VVCV) = (*VCV)

(VCV*) = (*VCV) = (VCV)

VCVC	3433	(VCVC) - (CVCV) = 25
CVCV	3408	
CVCC	1681	(CVCC) - (CCVC) = 25
CCVC	1656	
VCCV	1599	(CCVC) - (VCCV) = 57
CVVC	686	(CVVC) - (VVCV) = 167
VVCV	519	
VCVV	494	(VVCV) - (VCVV) = 25
VCCC	350	(VCCC) = (CCCV)
CCCV	350	(CCCV) - (CCVV) = 57
CCVV	293	(CCVV) - (VVCC) = 25
VVCC	268	(VVCC) - (VVVC) = 167
VVVC	101	
CVVV	101	(VVVC) = (CVVV)
CCCC	53	
VVVV	5	
Total :	14997	Différences

Il y a donc, entre les seize nombres du tableau, des relations nécessaires : on ne peut modifier l'un d'eux sans modifier aussi quelques autres.
On va montrer, sur cet exemple, qu'un tel schéma numérique comporte neuf degrés de liberté.

Pour commencer on notera que le nombre des (CCCC) et celui des (VVVV) est absolument libre. En effet, dans une séquence quelconque il est facile de modifier l'un ou l'autre de ces nombres sans toucher aux autres.

En second lieu, on a toujours (VCCC) = (CCCV) et (CVVV) = (VVVC).
En effet, si l'on rencontre, quelque part dans une séquence, le motif ... VCCC ..., il faut bien qu'il soit suivi, plus loin, d'un retour au signe V (n'oublions pas qu'il s'agit de séquence circulaire) ; et au premier retour on aura ... CCCV ...

Avec ces deux simplifications préalables, on peut décrire comme suit la forme générale à laquelle appartient le tableau numérique précédent

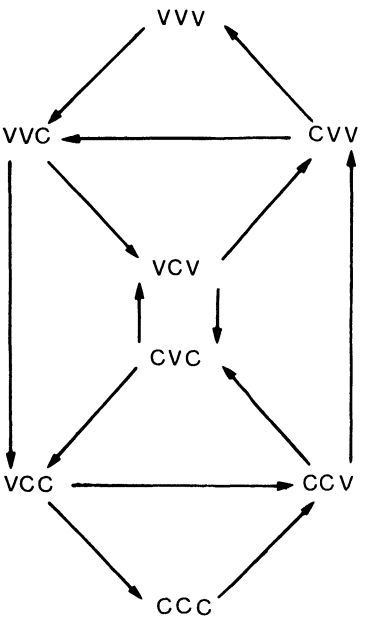
$$
\begin{aligned}
(VCVC) &= a + x \\
(CVCV) &= a \\
(CVCC) &= b + x + y \\
(CCVC) &= b + y \\
(VCCV) &= b \\
(CVVC) &= c + x + t \\
(VVCV) &= c + x \\
(VCVV) &= c \\
(VCCC) = (CCCV) &= d + x + y + t \\
(CCVV) &= d + x + t \\
(VVCC) &= d + t \\
(VVVC) = (CVVV) &= d
\end{aligned}
$$

Pour l'exemple choisi, on avait :

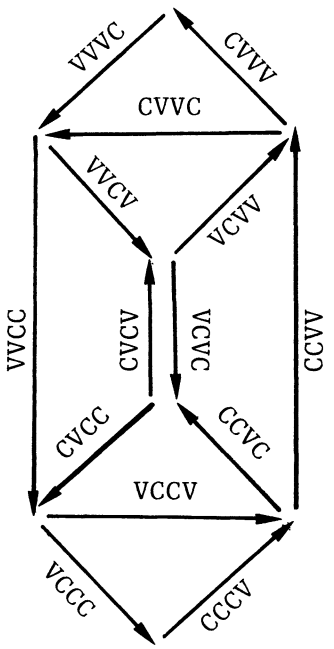
$$
\begin{aligned}
a &= 3408 & x &= 25 \\
b &= 1599 & y &= 57 \\
c &= 494 & t &= 167 \\
d &= 101
\end{aligned}
$$

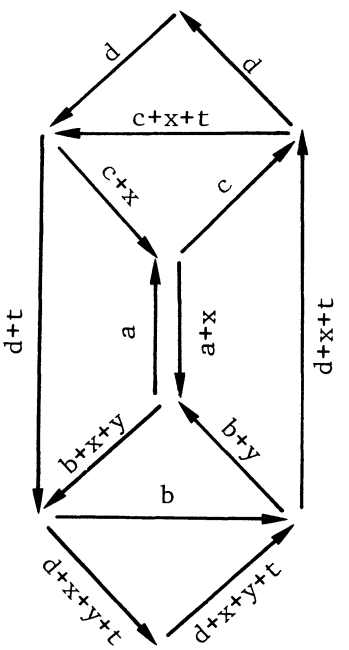
Je dis que ces sept nombres entiers peuvent être choisis comme on voudra.

Reprenons en effet le diagramme qui transcrit les règles d'enchaînement des triplets : VVC doit être suivi soit de VCV, soit de VCC, etc...

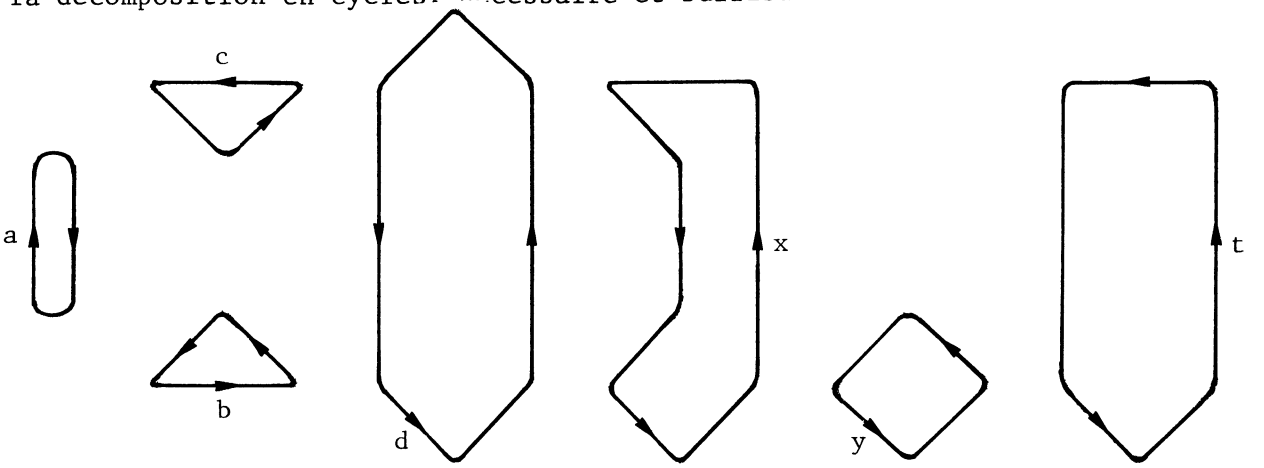


Et notons que sur *le même* diagramme on peut lire les règles d'enchaînement des quadruplets : il suffit de marquer, non plus les noeuds du réseau, mais les arêtes.
Toute séquence est alors une circulation dans un tel réseau.





En plaçant sur chaque arête la valeur numérique imposée, on voit apparaître la décomposition en cycles nécessaire et suffisante :



Or chaque cycle est aisément réalisable.
Il suffit alors de voir comment on peut les fondre en un seul, ce qui est facile.

On voit ainsi que les relations qui avaient été observées sur la statistique des quadruplets (prose de Pouchkine) sont nécessaires et qué ce sont les seules ; les voici :

(VCVC) - (CVCV) = (CVCC) - (CCVC) = (VVCV) - (VCVV) = (CCVV) - (VVCC)
(CCVC) - (VCCV) = (CCCV) - (CCVV)
(CVVC) - (VVCV) = (VVCC) - (VVVC)
(VCCC) = (CCCV)
(CVVV) = (VVVC)

Ce sont les *sept* conditions à remplir par les *seize* nombres pour qu'on puisse (d'un grand nombre de façons) réaliser une séquence conforme à cette statistique.

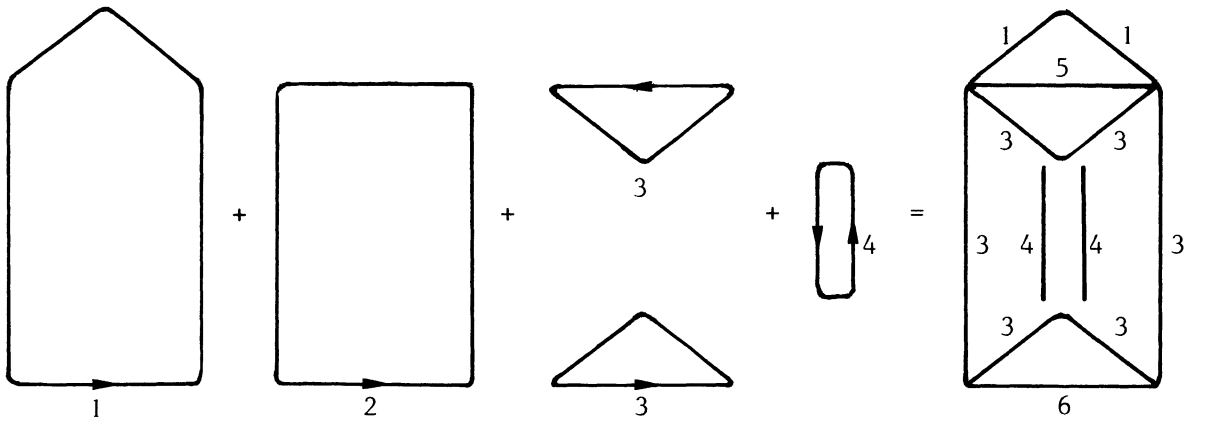
7. Bien entendu, dans l'exemple numérique qu'on vient d'analyser, les trois différences (les trois lignes de la liste ci-dessus, respectivement nommées x, y et t) étaient positives. Cela n'est pas nécessaire. Mais si l'un ou l'autre de ces nombres était négatif, la décomposition en cycles serait différente, évidemment, de celle qui a été dessinée ci-dessus. On ne peut s'étendre ici sur cette casuistique (qui peut cependant intéresser les applications phonologiques, car l'ordre des fréquences des divers quadruplets a probablement une signification intéressante).

Contentons-nous de reprendre le tout premier exemple fourni dans cette note (et qui transcrivait la séquence des voyelles et consonnes dans la première phrase de la dite note). On commencera par fermer la séquence en boucle, en opérant la confusion de la première et de la dernière lettre. Ce qui introduit deux quadruplets en sus : CVVC et VVCV. Le décompte est alors le suivant (on a fait figurer les différences) :

```
          (VCVC)  =  4
          (CVCV)  =  4  ..............................  x = 0

          (CVCC)  =  3
          (CCVC)  =  3  ..............................  x = 0
          (VCCV)  =  6                                   y = -3

          (CVVC)  =  5
          (VVCV)  =  3  ..............................  t = 2
          (VCVV)  =  3                                   x = 0

(VCCC)=(CCCV)  =  0
          (CCVV)  =  3  ..............................  y = -3
          (VVCC)  =  3                                   x = 0
(VVVC)=(CVVV)  =  1                                      t = 2
```

On peut souhaiter n'avoir que des paramètres positifs (éliminer y). C'est facile, il suffit de modifier l'ordre de présentation.

Voici d'ailleurs une décomposition en cycles (selon le schéma graphique déjà utilisé pour le texte russe) :



8. On peut généraliser. Pour les n-uplets, le réseau comporte 2^n arêtes et 2^{n-1} sommets.

Le nombre des relations à imposer est $(2^{n-1} - 1)$.

Le nombre des degrés de liberté (ou des cycles indépendants) est $(2^{n-1} + 1)$.

Les relations ont toutes la même forme : linéaires et homogènes à coefficients égaux à +1 ou -1.

En particulier, il existe toujours (au moins) une séquence circulaire présentant tous les arrangements de signes, c'est-à-dire tous les n-uplets possibles avec la *même* fréquence pour tous (c'est le problème dit "de Posthumus")*.

---

* Voir : *Cahiers Mathématiques*, III (collection Math. et Sci. de l'Homme, Paris, Gauthier-Villars et Mouton, 1970), pp.63-68.

et : *S.K. Stein*, Les mathématiques, ce monde que créa l'homme, Paris, Dunod 1967, pp.122-134.