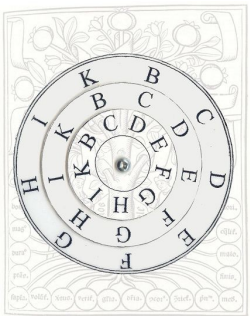
# Mathématiques et linguistique

par Pascal Kaeser, 2013



## 1. Quelques précurseurs

Au 13e siècle, Raymond Lulle, associant des mots à des lettres qu'il dispose sur 3 cercles concentriques mobiles, élabore par combinatoire un ensemble de questions philosophiques, par exemple: la bonté est-elle grande en ce qu'elle contient des choses différentes?

De Lulle et des pionniers de la cryptographie, Giordano Bruno reprend au 16e siècle l'idée des cercles concentriques. Il forme 150 bigrammes en collant successivement chacune des 5 voyelles à la droite de chacun des 30 caractères (23 latins, 7 hébreux et grecs) de l'alphabet qu'il a choisi. La série des bigrammes est répétée sur 5 cercles mobiles. Pour engendrer des phrases, Bruno fait correspondre 5 mots à une même lettre, selon sur quel cercle elle se trouve (1. personnage, 2. action, 3. insigne, 4. personnage, 5. circonstance). Le jeu combinatoire permet alors de produire des images comme: « une femme à cheval sur un taureau peigne ses cheveux en tenant un miroir dans sa main gauche, tandis qu'un adolescent avec un oiseau vert sur la main assiste à la scène ».

Georg Philipp Harsdörffer

SAGT DOCH: WAS IST DIE LIEB?

Ein Band vereinter Hertzen. Ein
überſüſſer Gifft. Ein angenemer Schmer-
tzen. Ein Pfeil der all- zeit trifft. Das Werck
ſo Menſchen mehret. Der Jugend friſcher Muht.
Ein Spiel das ſich verkehret. Cupido Flammen
Glut. Ein Laſt der leicht zu tragen. Ein angene-
mes Kind. Ein Trauren nach behagen. Ein
Strick der Freyer bindt. Ein blindt verfin-
ſtert Weſen. So manche gute Nacht.
Ein Buch das offt zu leſen. Der
Schönheit ſchneller Pracht. Ein
Marckt die Reu zu kauffen.
Ein kluger Unverſtand.
Ein Weg der machet
ſchnauffen. Ein
ſtets erneuter
Brand.

Au 17e siècle, le poète et savant touche-à-tout Georg Philipp Harsdörffer se laisse lui aussi séduire par les cercles concentriques mobiles. En plaçant 48 syllabes préfixes sur le 1er cercle, 60 chaînes de caractères sur le 2e, 12 caractères sur le 3e, 120 chaînes de caractères sur le 4e et 24 syllabes suffixes sur le 5e, il parvient à produire 99'532'800 mots allemands (dont bien sûr beaucoup n'existent pas).

Un hétérogramme est un mot formé de lettres différentes. Au 17e siècle, le savant jésuite suisse Paul Guldin calcule le nombre d'hétérogrammes possibles avec un alphabet de 23 lettres. Il s'agit de sommer, pour k variant de 1 à 23, les arrangements de k lettres parmi 23. On obtient environ 7E22. Guldin étudie en détail combien de livres il faudrait pour écrire tous ces mots et combien de bibliothèques pour contenir tous ces livres. Il démontre qu'avec des constructions cubiques de 432 pieds de côté, la surface terrestre ne suffirait pas.

Clavius, Mersenne et Leibniz ont envisagé des problèmes du même genre. Ce dernier se pose la question suivante: quel est le nombre maximal d'énoncés (vrais, faux, insensés) formulables avec un alphabet de 24 lettres? Si M est la taille maximale d'un énoncé, la réponse est la somme d'une suite géométrique de raison 24 et de premier terme 24, soit: (24^(M+1) – 24)/23. Leibniz prend pour M le nombre de caractères, à raison de 100'000 par jour, qu'un homme peut lire en 1000 ans (parce que l'alchimiste Artéphius aurait vécu aussi longtemps), ce qui donne environ: 3.65E10. Un petit calcul de logarithme nous permet alors de voir que le nombre maximal d'énoncés comporte plus de 50 milliards de chiffres.

Dans une œuvre de jeunesse, Leibniz envisage de hiérarchiser les idées. En combinant par 2 les idées simples et primitives d'une 1re classe, il obtient une 2e classe; en les combinant par 3, une 3e classe; etc. Il représente les termes de la 1re classe par des nombres et les combinaisons par des groupes de nombres. Quelques années plus tard, il améliore cette notation en associant des nombres

premiers aux termes de la 1re classe et des produits aux combinaisons. Ainsi, les termes de la 2e classe correspondent aux nombres qui ont 2 facteurs premiers, etc. La divisibilité devient alors un critère pour savoir si une idée de classe N entre dans la composition d'une idée de classe supérieure à N.

*
* *

2. Un peu de statistique

La notion de chaîne de Markov est née d'une étude linguistique. En 1913, Markov dégage cette idée d'une recherche statistique portant sur les 20'000 premières lettres de l'*Eugène Onéguine* de Pouchkine, en ne retenant que l'aspect voyelle-consonne. Il s'intéresse à des questions comme: quelle est la probabilité qu'une voyelle succède à une consonne?

Les fréquences des lettres, des bigrammes et des trigrammes font les délices des cryptanalystes. Les fréquences des phonèmes, des couples et des triplets de phonèmes intéressent les professionnels de l'audition, des troubles du langage, de la reconnaissance automatique de la parole, etc. D'après un ouvrage datant de 1990, les phonèmes les plus fréquents en français sont:
/a/ (8.1%), /r/ (6.9%), /l/ (6.8%), /é/ (6.5%), /s/ (5.8%), /i/ (5.6%).

La politique et la littérature sont deux cibles de choix pour la lexicométrie. L'étude des discours présidentiels montre que, par rapport à ses prédécesseurs depuis de Gaulle, Chirac abusait du présent de l'indicatif.
Le substantif le plus souvent associé à « vie » est « fois » chez Stendhal, « cœur » chez Flaubert, « mort » chez Proust.
En remplaçant chaque mot d'un texte par une lettre codant sa fonction grammaticale, une analyse factorielle des séquences de 3 codes successifs permet de dire qu'*Emile* (de Rousseau) est proche de *Marianne* (de Marivaux), mais éloigné de *Madame Bovary* ou du *cousin Pons*.

L'*Ulysse* de Joyce rendit Zipf célèbre, parce qu'il découvrit dans cette œuvre une relation entre la fréquence F de chaque mot et son rang R dans le classement par fréquences décroissantes. Plus précisément, le produit FR est constant. Par exemple, si le mot le plus fréquent apparaît 300 fois, le 2e mot le plus fréquent apparaîtra 150 fois, le 3e 100 fois, le 4e 75 fois, etc. C'est ce qu'on appelle la loi de Zipf. Trop simple pour être vraie, cette formule a été modifiée par Mandelbrot.

Des expériences le montrent: les scores de mémorisation d'un texte sont meilleurs quand les phrases et les mots sont de petite taille. Rudolph Flesh en a tiré le Reading Ease Level (plus connu maintenant sous la dénomination d'indice de lisibilité de Flesh). La formule est simple (trop pour être honnête): $F(x;y) = 206.835 – 1.015x – 84.6y$, où x est le nombre moyen de mots par phrase et y le nombre moyen de syllabes par mot.
Théoriquement, F(x;y) peut varier de moins l'infini à 206.835; en général, il se situe entre 0 et 100. Plus cet indice est élevé, plus le texte devrait être lisible (sauf bien sûr si l'auteur emploie des mots rares, des tournures extravagantes ou verse dans le charabia). Le Reading Ease Level du Reader's Digest vaut 65; Saint-Ex plane à 30; Proust chute à moins 10.
L'article 38a/699a de la législation du Connecticut stipule que l'indice de Flesh d'une police d'assurance ne doit pas descendre en dessous de 45. Il en va de même pour les formulaires de consentement éclairé qu'utilisent les hôpitaux.
La condition : $F(x;y) \geq 45$ donne: $x + 83.35y \leq 159.443$. Elle implique que y ne doit pas dépasser 1.901.

Un exemple: « Sandra lit dans l'âme de Pascal. Que voit-elle? La même chose que Pascal lit dans l'âme de Sandra. Des beautés à partager. » livre x = 25/4 = 6.25 et y = 35/25 = 1.4; donc son indice de Flesh vaut 82.

À vrai dire, cet indice est calibré pour la langue anglaise. Il est abusif de l'utiliser en français.

*
* *

3. Les cliques en sémantique

En 1998, Ploux et Victorri ont mené une étude sur la polysémie de « sec ». Ils ont construit un graphe dont les 63 sommets sont le mot « sec » et ses synonymes trouvés dans 7 dictionnaires. Chaque fois que deux de ces 63 mots sont synonymes, une arête les relie. Ce graphe n'est pas complet, car les synonymes de « sec » ne sont pas tous synonymes entre eux, par exemple « brusque » et « maigre » ne sont pas synonymes.

Une clique est un ensemble maximal de sommets tous reliés deux à deux. Le graphe des synonymes de « sec » comporte 94 cliques, par exemple {sec; fauché; pauvre} ou {sec; bref; brusque; tranchant}.

Les chercheurs définissent l'espace sémantique de « sec » par une projection à deux dimensions du nuage de points que les cliques forment dans l'espace multidimensionnel engendré par tous les mots considérés. Quelques précisions: on numérote les 63 mots; une clique devient un point à 63 coordonnées; chaque coordonnée vaut 1 quand sa position est le numéro d'un mot qu'elle contient; autrement, elle vaut 0; on munit cet espace de la distance du $\chi^2$; enfin, on projette les points sur un plan qui passe par le centre de gravité du nuage et qui déforme le moins possible la dispersion.

Cette méthode permet de visualiser les différents sens de « sec » et de les organiser, puisque la proximité sémantique correspond à la proximité géométrique. Six acceptions principales se dégagent alors: 1. qui manque d'eau; 2. maigre, décharné; 3. stérile, improductif; 4. qui manque de sensibilité; 5. bref, abrupt; 6. seul.

Et le plus beau, c'est que tout ce travail peut être effectué par un logiciel, en cinq sec!

*
* *

4. Formules et arbres pour représenter des structures syntaxiques

La notation utilisée en calcul des prédicats peut aider à dégager les relations présentes dans une phrase. Ainsi: « Pascal souhaite se rapprocher de Sandra » est décrit par la formule:

$$\mathbf{\exists R_1 \exists R_2 \ R_1(Pascal,R_2) \ \& \ R_2(Pascal,Sandra)},$$

c'est-à-dire: il existe deux relations $R_1$ (=souhaiter) et $R_2$ (=se rapprocher de) telles que $R_1$ unit Pascal à $R_2$ et $R_2$ unit Pascal à Sandra. Si l'on préfère:

$R_1$ dit: Pascal souhaite $R_2$ et $R_2$ dit: Pascal se rapproche de Sandra.

Notez que c'est Pascal qui énonce et qui analyse la phrase « Pascal souhaite se rapprocher de Sandra ». Or, la traduction qu'il en donne réalise son souhait, puisqu'à la fin de la formule le mot Pascal est tout proche du mot Sandra. Exercice: trouvez une formule qui rende compte de toute la complexité de ce phénomène.

La grammaire catégorielle de Bar-Hillel (1953) associe à chaque mot un ou plusieurs types (représentés par des formules) qui indiquent les classes possibles pour ses voisins de gauche et de droite. Je modifie un peu la notation d'origine. $X/Y_d$ signifie, sur le plan de la syntaxe, que la classe X peut être suivie de la classe Y; et, sur le plan de l'algèbre, qu'on obtient X en multipliant à droite

cette fraction par Y. La signification de $X/Y_g$ est analogue: il suffit de remplacer « suivie » par « précédée » et « droite » par « gauche » dans la phrase définissant $X/Y_d$.

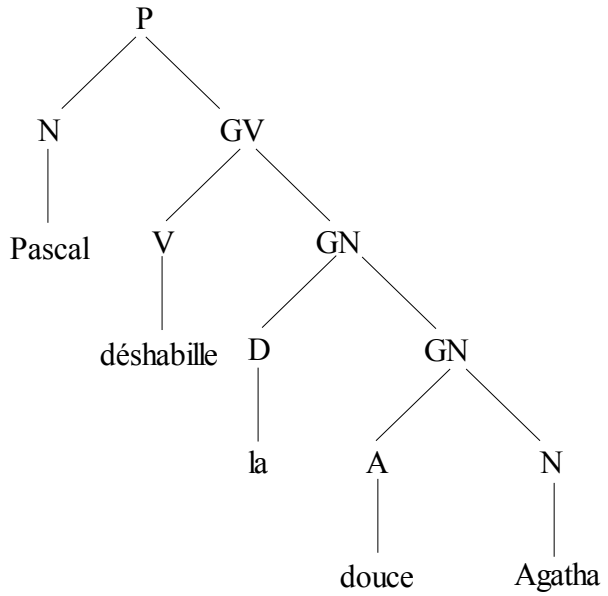
Exemple de type: $N/D_g$ signifie qu'un groupe nominal N peut être précédé d'un déterminant D (article, possessif, démonstratif, adjectif interrogatif, relatif ou indéfini, numéral).

Pour vérifier que « Pascal examine la mystérieuse Agatha » est bien une phrase P, on effectue le calcul suivant: $N\times P/(N_gN_d)\times D\times A\times N/(A_gD_g)$, où N désigne un groupe nominal, D un déterminant, A un adjectif qualificatif. Dans ce produit, la simplification des 2 premiers termes et des 2 derniers donne: $P/N_d\times D\times N/D_g$ qui se simplifie en $P/N_d\times N$, puis en P. Il s'agit donc bien d'une phrase.

Cette théorie présente une difficulté: choisir pour chaque mot le « bon type », celui qui, dans le contexte donné, conduit à la simplification maximale.

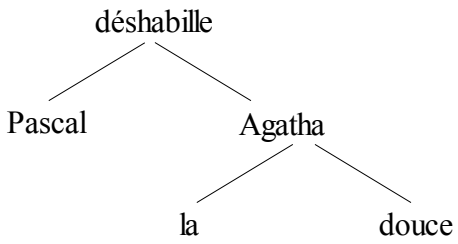
On trouve à l'origine de l'idée de grammaire catégorielle le philosophe Husserl (1913) et le logicien Lesniewski (1922). Les développements sont l'œuvre des linguistes Ajdukiewicz (1935), Bar-Hillel (1953), Lambek (1958), Steedman (1982). Mentionnons aussi Shaumyan (1965) qui opère un lien entre grammaire catégorielle et logique combinatoire de Curry (1958), un outil puissant voisin du λ-calcul de Church (1941). Desclés continue d'explorer cette piste.

Agatha raffole des arbres, ce qui me donne envie de suspendre un vêtement au bout de chaque branche d'un arbre syntagmatique de Chomsky (1957).




Symboles: P=Phrase, GN=Groupe nominal, GV=Groupe verbal, V=Verbe, D=Déterminant, A=Adjectif, N=Nom.

Pour décrire la structure d'une phrase, Tesnière, en 1959, imagine un autre genre d'arbre, le *stemma*, qui hiérachise les mots autour du verbe.



Un mathématicien d'aujourd'hui, Sylvain Kahane, invente le concept d'arbre à bulles pour étudier plus finement des phrases avec des propositions relatives.

*
* *

## 5. Grammaire générative

C'est au début des années soixante que Chomsky et les linguistes du M.I.T. élaborent l'ambitieuse théorie des grammaires génératives. Je me propose de la présenter au moyen d'un exemple. L'idée de base est celle de règle de réécriture.
X→Y signifie que X peut être remplacé par Y.
X→Y(Z) signifie que X peut être remplacé par YZ ou par Y, la parenthèse autour de Z rend cette lettre optionnelle.
X→{arbre/art/artichaut/univers} signifie que X peut être remplacé par n'importe lequel des mots mis entre accolades.

Une grammaire générative est alors un ensemble de règles de réécritures qui permettent de passer d'un symbole initial, noté ∑ et nommé axiome, à du texte, via des symboles intermédiaires, dits parfois « non terminaux ». Exemple:
∑→SVC
C→D(SVC)
D→NGN(D)
S→{Agatha/Sandra}
V→{m'enseigne}
G→{de}
N→{la nature/la beauté/l'amour}

Voici une dérivation possible (j'indique en gras la lettre que je remplace à chaque étape):
∑→SV**C**→SV**D**SVC→SVNGNDSV**C**→SVNGN**D**SVD→SVNGNNGN**D**SVD→
SVNGNNGNNGNSV**D**→SVNGNNGNNGNSVNGN**D**→SVNGNNGNNGNSVNGNNGN**D**→
SVNGNNGNNGNSVNGNNGNNGN→ Agatha m'enseigne la beauté de la nature, l'amour de la beauté, la nature de la nature. Sandra m'enseigne l'amour de la nature, la nature de l'amour, la beauté de l'amour.

*
* *

## 6. Vecteurs booléens et cas grammaticaux

Dans les années soixante, Dénes Varga eut l'idée d'associer à chaque mot un vecteur booléen qui rende compte de ses déclinaisons possibles.

Il y a 12 cas en russe: 1. nominatif singulier, 2. accusatif singulier, 3. génitif singulier, 4. datif singulier, 5. instrumental singulier, 6. prépositionnel singulier, 7. à 12. les mêmes au pluriel.

La préposition **с** ne peut intervenir que dans 6 cas: l'accusatif, le génitif, l'instrumental, chacun au singulier ou au pluriel. Représentons-la par un vecteur où nous plaçons 1 aux positions de ces 6 cas et 0 ailleurs: (0,1,1,0,1,0,0,1,1,0,1,0). De même le mot **нашей** admet 4 cas: génitif, datif, instrumental, prépositionnel, chacun au singulier exclusivement. Nous obtenons le vecteur: (0,0,1,1,1,1,0,0,0,0,0,0). Dans quels cas les mots **с** et **нашей** sont-ils compatibles? Il suffit d'effectuer le produit booléen terme à terme. Le vecteur de **с нашей** est (0,0,1,0,1,0,0,0,0,0,0,0), donc la réponse est: au génitif singulier et à l'instrumental singulier.

*
* *

## 7. Aperçu de théories plus pointues

Dans les années soixante (décidément une époque glorieuse pour la linguistique mathématique), Marcus s'attaque à la phonologie avec des espaces métriques, des demi-groupes, des algèbres de Boole, des graphes. Par exemple, la notion de nombre cyclomatique (=nombre maximal de cycles indépendants dans un graphe) donne un éclairage sur une question comme: dans quelle mesure la connaissance des extrémités d'un groupe consonantique situé au début d'un mot, ou la connaissance de la consonne initiale et de la voyelle qui vient immédiatement après le groupe, détermine-t-elle la connaissance du groupe tout entier?

On trouve chez Desclés une approche topologique, via l'algèbre de Kuratowski, de certaines situations sémantiques.

La traduction automatique nécessite de pouvoir à la fois modéliser le sens et la syntaxe. Le pari est encore loin d'être gagné, mais quel moteur pour la recherche!

La linguistique ne se contente pas de recourir à des structures mathématiques connues. Elle en invente de nouvelles.

*
* *

## Quelques références

Mes sources sont trop nombreuses pour toutes les citer. Voici une sélection:

Bar-Hillel, Yehoshua, A quasi-arithmetical notation for syntactic description, Language, 29(1), 1953
Biskri, Ismaïl, La grammaire catégorielle combinatoire applicative dans le cadre des grammaires applicatives et cognitives, Thèse de doctorat, LaLIC, Paris-Sorbonne, 1995
Brunet, Etienne, Fréquences et séquences. Mise en œuvre dans Hyperbase, Laboratoire BCL, 2007
Desclés, Jean-Pierre, Formes opératoires et topologiques en linguistique, Mathematics and Social Sciences 193(1), 2011
Eco, Umberto, La recherche de la langue parfaite, Seuil, 1994
Fayard, Luc, Comment nous lisons, doc. sur Internet, 2005
Kahane, Sylvain, La modélisation mathématique des langues naturelles, conférence, UTLS, 2002
Ploux, Sabine et Victorri Bernard, Construction d'espaces sémantiques à l'aide de dictionnaires informatisés des synonymes, TAL, 39(1), 1998
Solomon, Marcus, Introduction mathématique à la linguistique structurale, Dunod, 1967
Venant, Fabienne, Géométriser le sens, LaTTiCe-ENS, 2004