MOT ET IMAGE : LES MÊMES LOIS STATISTIQUES

Gilberte PÉROT

Nous publions ici un article de Madame G. PÉROT *qui, pendant de longues années, a été maître-assistant à l'Institut de Mathématiques de Strasbourg. Cet article a été publié en* 1972 *dans la revue 'Communication et Langage'. Il porte sur la loi de* ZIPF*, bien connue des linguistes, concernant les fréquences d'apparition des mots dans une langue naturelle. On sait que cette loi traduit en termes statistiques un phénomène d'économie du locuteur. Madame G.* PÉROT*, avec son intuition coutumière, avait pensé que cette loi ne s'appliquait pas seulement aux messages linguistiques, mais également aux messages iconiques (comportant des images); elle nous fait part, dans cet article, des résultats auxquels elle a abouti.*

Pour une étude approfondie de la loi de ZIPF *dans le cas classique on pourra consulter :*
A. FUCHS *: La théorie de l'information et la linguistique 'Gazette des Mathématiciens' (S.M.F.) n° 17 (juil.* 1981*).*

N'en déplaise aux iconoclastes (ils furent nombreux au cours des siècles, particulièrement chez les Hébreux, les Arabes, les Byzantins, les Protestants, les Bourbakistes), les images ont triomphé. Actuellement, le message iconique concurrence le message linguistique, et il a pour lui deux atouts. D'une part, il transmet plus rapidement l'information; pensons à un schéma de montage, à une formule développée, à un film présentant une opération chirurgicale ou comparant des espèces en botanique (le cinéma a suscité "un intérêt que cette science n'avait jamais réussi à éveiller auparavant"). D'autre part, la linéarité du langage, mise en évidence par SAUSSURE, s'adapte parfaitement aux "longues chaînes de raisons, toutes simples et faciles", mais la souplesse du message iconique convient beaucoup mieux à la transcription des structures complexes. Elle permet d'arriver, plus économiquement, à l'intelligibilité des réseaux comportant circuits, embranchements, feed-back, circulation d'un flux dont il faut préciser la valeur en différents points : pensons à un organigramme, à un graphe, à un sociogramme. Verrons-nous bientôt le jour où le langage iconique deviendra aussi familier que la langue maternelle? Cette question relève de la science-fiction, mais, en tout cas, je signale un fait qui incite à la réflexion : les nouveaux programmes de mathématique, pour les classes de sixième, prescrivent l'utilisation des schémas sagittaux et cartésiens. L'intérêt est évident, si l'on songe à la mathématisation des situations effectives et à l'utilisation des ordinateurs. La cybernétique a mis l'accent sur des modes nouveaux d'intelligibilité. La faveur croissante des schémas nous invite à nous demander si l'ensemble des représentations graphiques et des images que nous échangeons usuellement à des fins sémantiques possède ou non des structures comparables à celles d'une langue. Connaît-il les mêmes agencements, suit-il les mêmes lois?

La loi de ZIPF est valable pour les images

L'objet de cette étude est de poser le problème, de le circonscrire, et, dans les limites fixées, de présenter cinq corpus différents prouvant que les morphèmes (formes simples) vérifient, exactement comme le font les mots d'une langue naturelle, la loi de ZIPF-MANDELBROT.

Le flot des messages iconiques étant extrêmement riche et extrêmement hétérogène, il semble nécessaire de restreindre le sujet dès le départ. Le langage ordinaire sous-tend toujours l'information iconique, soit qu'il n'apparaisse qu'en filigrane, soit au contraire qu'il soit explicitement présent, et joue un rôle manifeste. Nous avons, dans ce dernier cas, des messages hybrides, où une partie de la transmission passe par le canal des images, l'autre par la parole ou l'écriture : film parlant ou sous-titré, télévision, histoires en images accompagnées de légendes (1). Nous excluons de notre étude ces "messages hybrides" ; si, par exemple, des inscriptions figurent dans un dessin, nous n'admettons ce dessin que dans la mesure où chiffres, lettres et mots n'ont d'autres fonctions que d'identifier des éléments de l'assemblage ou de préciser la valeur numérique d'une application, ou encore d'expliciter un codage. A plus forte raison les agencements de formes destinés à reconstituer des lettres, et un message à déchiffrer par la lecture, ne sont pas retenus : les films réalisés en composant des petits carrés par le truchement d'un ordinateur (Skopo) contiennent souvent des séquences de ce type.

Champ et limite de cette étude

Le champ où s'inscrit notre étude est, en fait, très étroit. Pour les messages iconiques, aux impacts émotionnels importants et souvent imprévus, aux connotations multiples, l'écart entre signifiant et signifié est bien plus considérable qu'en linguistique. Comment se délivrer de l'ambiguïté fondamentale des significations, dans un domaine où les modèles mêmes de la perception restent à étudier (2) ? L'idéal serait de pouvoir considérer le signifiant comme élément d'un ensemble E, le signifié comme élément d'un ensemble F, et la signification comme une relation fonctionnelle de E dans F, mais évidemment cela n'est jamais réalisé. Nous avons essayé de nous rapprocher de ce cas limite en nous plaçant aux confins mêmes des codages. C'est pourquoi nous nous sommes bornés à vingt formes très simples (précisées ultérieurement) ; c'est pourquoi aussi nos corpus sont les illustrations de livres techniques. Cela met très fortement l'accent sur le message sémantique, alors que nous savons bien que tout canal convoie simultanément information esthétique et information sémantique. Il suffit d'ailleurs de présenter plusieurs schémas logiquement équivalents pour qu'aussitôt s'impose la prédilection des sujets pour les figures régulières, ou présentant un axe de symétrie. Notre étude, volontairement placée dans le cadre le plus simple et le plus favorable, doit donc être complétée

(1) Cf. Enrico FULCHIGNONI : *La Civilisation de l'image* (Paris, N.R.F., 1967) ; Abraham MOLES : *Théorie de l'information et perception esthétique* (Paris, Flammarion, 1958) ; Abraham MOLES : *Sociodynamique de la culture* (Paris, Mouton, 1967).
(2) Cf. E. VURPILLOT : "Quelques théories et modèles de la perception", in '*Bulletin de Psychologie*' (15 avril 1967).

par un élargissement progressif du champ et des corpus. Nous sommes en train de dépouiller un corpus et songeons à des corpus de peinture abstraite (en particulier, œuvres de MALEVICH). Nous étudions aussi, en collaboration avec J.-Cl. SIDLER, les projections observées sur des figures comportant une structuration mathématique évidente. Dans une optique beaucoup plus large, si on veut mettre l'accent sur l'information esthétique, on peut consulter Hermann WEYL et Erwin PANOFSKY (3).

Le concept d'image étant ambigu, nous précisons enfin que nous nous limitons à des figures pouvant être dessinées sur une feuille de papier. Les éléments de forme auraient donc dû être appelés graphèmes. Nous avons préféré "morphèmes", car les lettres de l'alphabet sont des graphèmes, et nous voulions les exclure (sous les restrictions indiquées précédemment). Les morphèmes nous sont apparus, puisqu'ils suivent la loi de ZIPF, analogues aux mots. Ils peuvent être obtenus par une suite de tracés de segments et d'arcs, ces arcs pouvant eux-mêmes être obtenus par raccordements d'arcs de cercle (théorie de cercle osculateur). Le coût du tracé d'un morphème correspond peut-être à la longueur d'un mot en linguistique, et demanderait à être étudié. Pour ces "mots" que sont les morphèmes, nous ne disposons pas de tables de fréquence, mais les cinq corpus étudiés ont donné pratiquement les mêmes morphèmes les plus usités. Nous avons complété, en utilisant d'autres sources, à 20 morphèmes, qui seront considérés comme mots du vocabulaire de base. Notre étude ne porte que sur ces vingt morphèmes, rangés ici par analogie de tracé, et non par ordre de fréquence :

---

point
segment
triangle, carré, rectangle, losange, parallélogramme, trapèze, hexagone, octogone, chevron
ligne brisée
angle
arc
arc comportant plusieurs boucles identiques (cf. solénoïde en physique)
flèche
cercle
ellipse
"patate" utile pour les diagrammes de VENN ou les "cercles" d'EULER)
"terminal", c'est-à-dire rectangle dont les petits côtés sont remplacés par des demi-cercles : on l'appelle "terminal" en informatique.

Les limites de cette étude étant ainsi précisées, nous allons, dans une deuxième partie, rappeler la loi de ZIPF-MANDELBROT, puis montrer qu'elle s'applique aux morphèmes des corpus que nous avons considérés.

"La langue est un système dont tous les éléments sont solidaires, et où la valeur de l'un est solidaire de la présence des autres." Cette phrase de de SAUSSURE est une phrase clé de la linguistique. Elle nous dit, par exemple, que le nombre

(3) Hermann WEYL : *Symétrie et mathématique moderne* (Paris, Flammarion, 1964) ; Erwin PANOFSKY : *Essai d'iconologie* (Paris, N.F.R., 1967).

n d'apparitions du mots "*the*", dans un texte anglais, et le nombre N de mots du texte ne sont pas des quantités indépendantes. Si on utilise des textes longs, des livres, le rapport n/N est toujours 0,09 : c'est la fréquence du mot "*the*". De même, la fréquence du mot "*of*" est 0,05. On peut, plus généralement, calculer la fréquence d'un élément quelconque : ainsi, la fréquence de la lettre "*E*" est 0,17 en français, 0,10 en anglais.

La loi de ZIPF

ZIPF détermina systématiquement la fréquence des mots en anglais, puisque ses dénombrements portent sur 8 727 mots différents. Ces mots peuvent donc être rangés par ordre de fréquence décroissante. Le mot qui a la plus grande fréquence, "*the*", a pour rang 1, puis vient "*of*", fréquence 0,05, rang 2, etc ... ; "*quality*", fréquence 0,00011 a pour rang 1000 ... Nous désignerons par f_r la fréquence du mot de rang r.

La loi de ZIPF, énoncée en 1946, précise que *le produit de la fréquence d'un mot par son rang est constant.*

$$f_r \times r = k.$$

Pour l'anglais, la constante k est voisine de 0,10.
1) Les axes étant $x'ox$, $y'oy$, posons

$$x = r, \quad y = f_r, \quad xy = k.$$

Le graphe est une hyperbole équilatère.
2) Il est préférable d'utiliser du papier logarithmique pour que la représentation graphique donne une droite

$$\log x + \log y = \log k$$

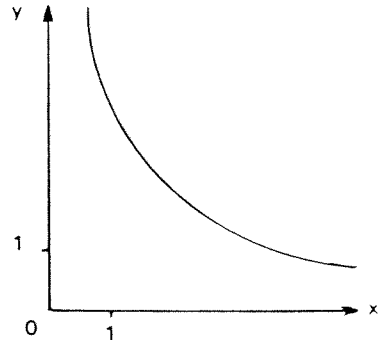
ou en posant $X = \log x, Y = \log y, K = \log k$

$$X + Y = K.$$

Les points du graphe sont alignés sur une droite de pente -1 : pour vérifier la loi de ZIPF, il suffira de vérifier cet alignement : c'est ce que nous ferons ultérieurement pour les morphèmes (4).

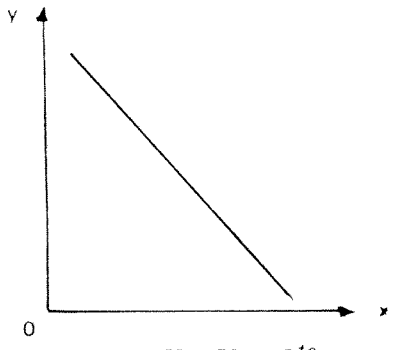
---

(4) Signalons l'apport de MANDELBROT. Comme toute loi empirique, la loi de ZIPF n'est qu'approximative : elle n'est bien vérifiée que par les mots de fréquence moyenne. Sur papier logarithmique, les points représentant ces mots sont alignés, mais les points représentant les premiers mots (plus grandes fréquences) et les derniers mots (plus petites fréquences) se répartissent respectivement sur 2 arcs de courbe. Pour rendre compte de ces faits, il faut remplacer, comme l'a indiqué MANDELBROT, la première expression donnée : $f_r \times r = k$ par $f_r(r+\beta)^\varphi = c$ qui exprime la loi de ZIPF-MANDELBROT : φ, β et c sont des constantes positives et φ est supérieur à l'unité.



$xy = C^{te}$

Courbe hyperbole équilatère



$X + Y = C^{te}$

La même courbe dans une représentation logarithmique (droite de pente -1)

Critères de choix des images

Passons maintenant de la linguistique au langage iconique. Voici les idées préconçues qui nous ont guidés. Dans certains domaines, les illustrations visent plus à être des compositions non figuratives que des copies fidèles de la réalité. Nous entendons par là que les images visent alors moins à reproduire les caractéristiques d'un objet qu'à mettre en relief l'agencement de divers éléments. Plus que les choses représentées importent les assemblages, c'est-à-dire finalement les structures. Parmi toutes les images, ce sont les plus intelligibles; il fallait donc commencer par elles l'étude du langage iconique, et pour cela choisir les corpus en conséquence : si l'on s'intéresse à l'information esthétique, s'en référer à la peinture abstraite (5), si on se limite à l'information sémantique, selon notre propos dans cet article, utiliser les illustrations de livres de mathématique, de physique ou d'informatique. Nous pensions que, dans ces derniers, les vingt morphèmes que nous avons précisés plus haut devaient être de beaucoup les plus fréquents. Ils s'imposaient par l'économie de tracé (segment, flèche, cercle, rectangle), par la facilité de raccordement (hexagone, terminal) et par la "raison de symétrie" (cercle, polygones réguliers).

D'autre part, le souci de lisibilité implique un codage : comment noter un rhéostat en physique, une sortie de résultats dans un organigramme? Les codes sont évidemment arbitraires. Mais, pratiquement, les questions d'économie jouent, l'environnement socioculturel est impliqué. L'utilisation des divers morphèmes, dans un code, se rode à l'usage et se rode pour arriver à une transmission plus efficace, de rendement meilleur. Ainsi, vraisemblablement, si des images sont utilisées à des fins sémantiques par de nombreux techniciens, l'ensemble doit se modifier, s'adapter progressivement pour devenir langue véhiculaire commune, pour devenir langage : il finit sans doute par refléter l'organisation linguistique. Nous devions donc, si cette opinion est correcte, vérifier que les morphèmes suivent la loi de ZIPF, et qu'ils la suivent mieux que les corpus tirés des mathématiques et de la physique que pour l'informatique, où le temps de rodage a été moindre. C'est bien ce qu'ont confirmé les résultats.

(5) Cf. le tableau "Huit rectangles rouges", de MALEVICH.

Choix des images

Le premier corpus choisi est constitué par les illustrations de l'ouvrage :
1) GALION : "*Mathématiques, classes de 6*^{e} " (O.C.D.L. Hatier, 1967).
Nous avons retenu ce livre parce qu'il concrétise l'effort d'un groupe très dynamique de professeurs lyonnais pour introduire la mathématique nouvelle au lycée : audacieux par son texte, il l'est aussi par une conception neuve de l'illustration et connaît un grand succès.

Le deuxième corpus a été choisi par analogie :
2) BRÉDIF : "*Mathématiques, classes de 6*^{e}" (Hachette, 1969).

Les corpus tirés de la physique correspondent simplement à des ouvrages récents et se vendent bien. C'est le libraire qui nous les a conseillés. D'autres corpus auraient donc, sans doute, donné une vérification expérimentale aussi bonne.
3) GUILHIEN : "*Electronique, tome 1*" (P.U.F.)
(Tubes électroniques à vide amplificateurs).
4) KASSATINE et PÉRÉKALINE : "*Cours d'électrotechnique*" (M.I.R.).
Le dernier corpus est constitué par les illustrations d'un livre de **programmation** :
5) RALSTON et WILF : "*Méthodes mathématiques pour calculateurs*".
La vérification de la loi de ZIPF ayant été satisfaisante dans ces 5 corpus, les premiers essayés, nous n'avons pas fait d'autres tentatives, sauf celle notée ci-dessous.
6) *Corpus constitué de 14 numéros de "L'Usine nouvelle".*

Résultats de l'étude

1. **Galion** : « Mathématiques 6e » (O.C.D.L., Diffusion Hatier, Paris, 1968, 93 pages).

Nombres de morphèmes : N = 3 700.
Morphèmes les plus fréquents : 3 657.
Autres morphèmes : 40.

Morphèmes	*Rang* r	*Nombre* n	*Fréquence* f_r	*Fréquence × rang* $f_r \times r$
Cercles	1	1 245	0,33	0,33
Flèches	2	645	0,17	0,34
Rectangles	3	452	0,12	0,36
Lignes courbes	4	325	0,087	0,35
Segments	5	269	0,072	0,36
Angles	6	213	0,057	0,34
Triangles	7	189	0,051	0,35
Carrés	8	171	0,046	0,36
Symboles :				
Astérisques, etc.	9	148	0,040	0,35
Autres morphèmes	10	40		

RÉSULTAT : le produit $f_r \times r$ *est constant, voisin de 0,35.*

2. Brédif : « Mathématiques, classe de 6e » (Hachette, juin 1969, 205 pages).

Nombre de morphèmes N = 3 108.
Morphèmes les plus fréquents : 2 908.
Autres morphèmes : 200.

Morphèmes	*Rang* r	*Nombre* n	*Fréquence* f_r	*Fréquence × rang* $f_r \times r$
Ronds	1	1 084	0,34	0,34
Flèches	2	574	0,18	0,36
Segments	3	384	0,12	0,36
Patates	4	279	0,089	0,35
Rectangles	5	228	0,073	0,35
Triangles	6	193	0,062	0,36
Carrés	7	166	0,053	0,37
Autres morphèmes	8	200		

RÉSULTAT : le produit $f_r \times r$ est constant, voisin de 0,35.

3. Guilhien : « Électronique, tome 1 (tubes électroniques à vide amplificateurs) » (P.U.F., 510 pages).

Nombres de morphèmes : N = 6 216.
Morphèmes les plus fréquents : 5 716.
Autres morphèmes : 500.

Morphèmes	*Rang* r	*Nombre* n	*Fréquence* f_r	*Fréquence × rang* $f_r \times r$
Segments	1	2 249	0,36	0,36
Fleches	2	1 380	0,22	0,44
Lignes courbes	3	874	0,14	0,42
Rectangles	4	682	0,109	0,43
Rhéostats	5	532	0,085	0,42
Autres morphèmes	6	500		

RÉSULTAT : le produit $f_r \times r$ est constant, voisin de 0,42.

4. Kassatine et Pérékaline : « Cours d'électrotechnique » (Éditions M.I.R., Moscou, 679 pages).

Nombre de morphèmes : N = 6 690.
Morphèmes les plus fréquents : 6 238.
Autres morphèmes : 452.

Morphèmes	*Rang* r	*Nombre* n	*Fréquence* f_r	*Fréquence × rang* $f_r \times r$
Flèches	1	2 730	0,40	0,40
Ronds	2	1 390	0,207	0,41
Rectangles	3	892	0,133	0,40
Lignes courbes	4	685	0,102	0,41
Segments	5	541	0,081	0,40
Autres morphèmes	6	452		

RÉSULTAT : le produit $f_r \times r$ est constant, voisin de 0,41.

5. Ralston et Wilf : « Méthodes mathématiques pour calculateurs arithmétiques » (1965, 477 pages).

Nombre total de morphèmes : N = 3 370.
Morphèmes les plus fréquents : 2 970.
Autres mosphèmes : 400.

Morphèmes	*Rang* r	*Nombre* n	*Fréquence* f_r	*Fréquence × rang* $f_r \times r$
Flèches	1	1 470	0,43	0,43
Rectangles	2	703	0,208	0,42
Segments	3	462	0,137	0,41
Ronds	4	335	0,099	0,40
Autres morphèmes	5	400		

RÉSULTAT : le produit $f_r \times r$ est constant, voisin de 0,42.

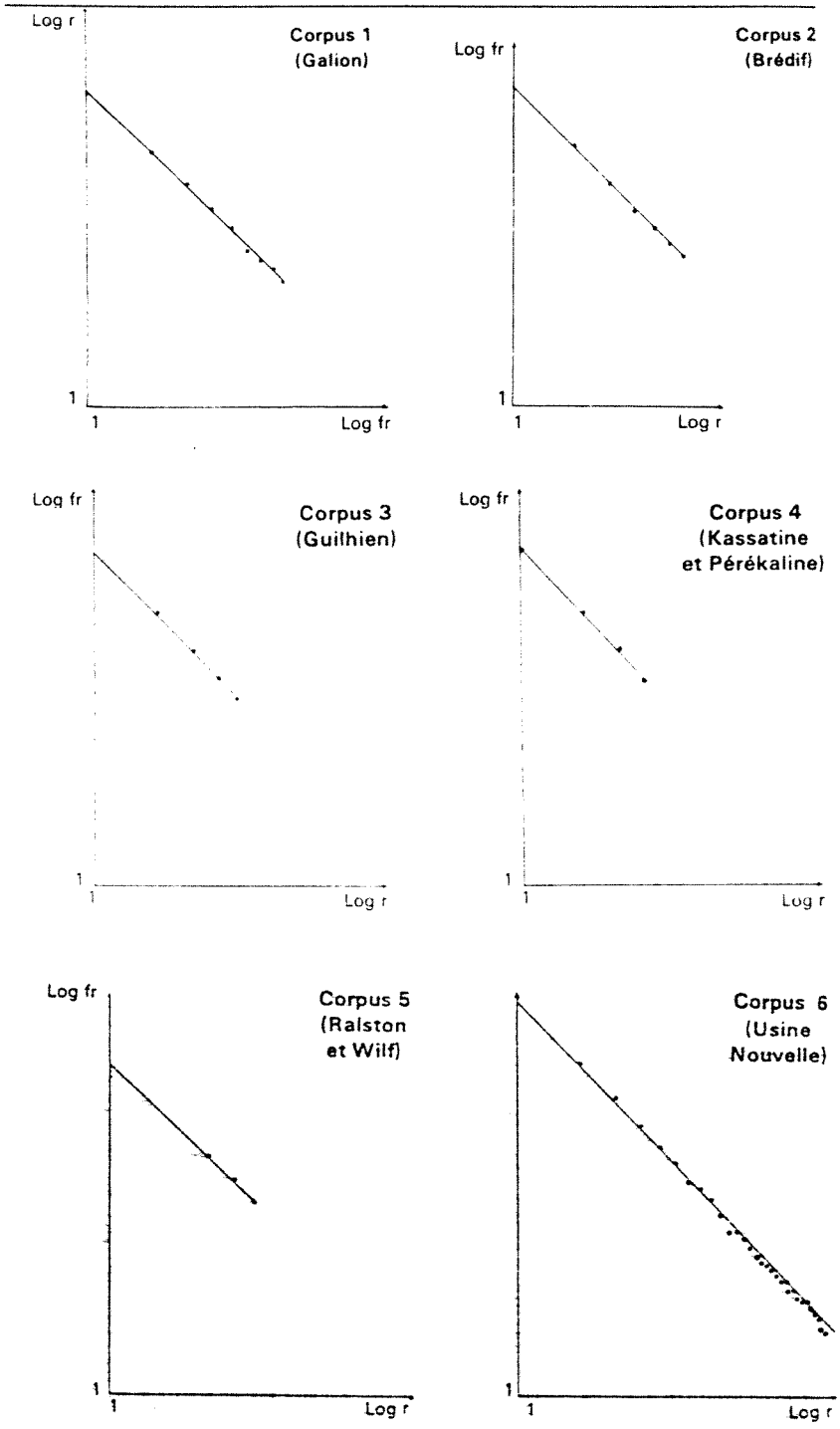
6. Usine Nouvelle : 14 numéros.

Nombre total de morphèmes : N = 3 200.
Morphèmes les plus fréquents : 3 146.
Autres morphèmes : 54.

Morphèmes	*Rang* r	*Nombre* n	*Fréquence* f_r	*Fréquence × rang* $f_r \times r$
Rectangles	1	800	0,25	0,25
Cercles	2	395	0,12	0,24
Flèches	3	270	0,084	0,25
Ellipses	4	198	0,061	0,24
Carrés	5	158	0,049	0,24
Roues dentées	6	134	0,041	0,25
Losanges	7	106	0,033	0,23
Terminal	8	101	0,031	0,25
Hexagones	9	90	0,028	0,25
Triangles	10	76	0,023	0,23
Chevrons	11	62	0,019	0,22
Polygones étoilés	12	62	0,019	0,23
Parallélogrammes	13	57	0,017	0,23
Trapèzes	14	52	0,016	0,22
Flammes	15	47	0,014	0,22
Ellipses « paraphes »	16	45	0,014	0,22
Sablonnés	17	43	0,013	0,22
Rectangles coins arrondis	18	41	0,012	0,23
Pentagones	19	38	0,011	0,22
Patates	20	36	0,011	0,22
Carrés sur pointe	21	36	0,011	0,23
Arcs « enroulés »	22	32	0,010	0,22
Demi-cercles	23	32	0,010	0,23
Porte-clefs	24	30	0,0093	0,22
Demi-terminal	25	28	0,0087	0,21
Accolades	26	28	0,0087	0,22
Gouttes et bulles	27	27	0,0084	0,22
Triangles incurvés	28	26	0,0081	0,22
Quadrilatères incurvés	29	25	0,0078	0,22
Ondes	30	24	0,0075	0,22
Arcs d'hyperbole	31	21	0,0065	0,20
Octogones	32	20	0,062	0,20

AUTRES MORPHÈMES A SIGNALER
LIGNES DENTS DE SCIE ou bords analogues (18), polygones concaves (10), polygones de 7 côtés (7), cœur (5), main (2), œil (3), écusson (9).
VALEUR DE LA CONSTANTE : 0,23

Courbes correspondantes



Pour chaque corpus, nous donnons d'abord le tableau de dépouillement, les morphèmes étant classés suivant l'ordre de fréquence décroissante, puis le graphe obtenu sur papier logarithmique. L'alignement des points représentatifs concrétise la loi de ZIPF. Il se réalise très bien pour les cinq premiers corpus, tandis que l'accord est moins bon pour le sixième, ce qui est naturel puisque le nombre total de morphèmes y est petit. C'est le seul cas qui nécessite une correction et l'application de la loi de ZIPF-MANDELBROT.

Nous pouvons donc conclure que, dans les corpus analysés, le langage iconique vérifie la loi de ZIPF, avec une précision meilleure qu'en linguistique, puisque, sur les "textes longs" il a été inutile de corriger en appliquant la loi de ZIPF-MANDELBROT (6).

La loi de ZIPF traduit d'ailleurs, comme on le démontre en théorie de l'information, une tendance des langages à se modifier dans le sens d'une transmission plus efficace; il est donc assez naturel de la retrouver dans des lexiques différents. Mais, si le produit de la fréquence par le rang est constant, pour les morphèmes comme pour les mots, il faut remarquer que la valeur de la constante n'est pas la même en iconologie et en linguistique : de l'ordre de 0,4 dans le 1er cas (nous avons trouvé 0,35; 0,35; 0,42; 0,41; 0,42; 0,23), de l'ordre de 0,1 dans le second cas; cela tient à la pauvreté du vocabulaire iconique considéré, par rapport à celui de la langue naturelle.

Difficultés d'élargir le champ de l'analyse

D'autre part, la précision des résultats ne doit pas faire illusion : nous avons insisté sur le fait que nous nous limitions à un champ très étroit et particulièrement favorable. L'exploration d'un champ plus large est délicate. La délimitation d'un mot, en français, pose relativement peu de problèmes, sauf le cas de l'apostrophe, du trait d'union, des mots composés (7). Au contraire, la délimitation d'un morphème est très complexe : les morphèmes composés sont légion (par exemple : résistance fixe codée rectangle, résistance variable codée rectangle traversé d'une flèche), et certains morphèmes sont ambigus (par exemple : arc). D'autre part, la "dimension d'iconicité" est très variable (8). Les dessins s'échelonnent de la représentation la plus fidèle du réel à la représentation symbolique, du figuratif au non-figuratif. L'identification d'un morphème est donc peu nette, sauf dans le cas extrême où, la représentation étant symbolique, l'identification se fait par codage. Il semble intéressant de tenir compte des structures mathématiques sous-jacentes (ensembliste, topologique, linéaire, métrique) pour préciser la notion d'iconicité, étudier les fonctions des morphèmes et décrire ces assemblages.

---

(6) Voir note 7. Cf. Charles MULLER : *Initiation à la mathématique linguistique* (Larousse) où figure page 168 un tableau donnant le produit de la fréquence par le rang, que l'on peut utilement comparer aux tableaux que nous avons donnés pour nos corpus.
(7) Cf. l'ouvrage de Charles MULLER cité, pages 145 à 151.
(8) Cf. A MOLES : "Théorie informationnelle du schéma" in revue *'Schéma et schématisation'* (n° 1, 1968).