

Comment nous lisons

Pour ceux que ça intéresse, voici un petit résumé de mon cours de Dauphine « Maîtriser son information » sur la partie lecture.

Savoir comment on lit permet ensuite de mieux comprendre comment il faut écrire pour être lu et compris. Je reviendrai plus tard sur cette partie écriture.

La plupart des informations que je traite ici viennent des enseignements des écoles de journalisme, notamment le CFPJ à Paris et l'ESJ à Lille.

Le lecteur parcourt un journal selon différents niveaux de lecture. Il consacre en moyenne 15 à 20 minutes à la lecture d'un quotidien. Sachant qu'il peut lire 12 000 mots à l'heure, il va donc en lire effectivement 3 000 à 4 000 mots, soit moins de 10% de la surface du quotidien. Appliquez cette règle au contenu des blogs et vous désespérez immédiatement tous les auteurs ! Alors, restons positifs et concrets !

## a) Les trois niveaux de mémoire

### • La mémoire de travail

C'est celle qui est d'abord mise en œuvre dans le processus de lecture. Elle a une capacité maximale d'affichage de 7 à 10 mots qui sont comme autant de spots de l'œil et qui peuvent être retenus pendant plusieurs secondes. Le processus de lecture se fait par écrêtage : on s'arrête sur des groupes de syllabes, ceux qui sont nécessaires pour identifier le mot et on en oublie d'autres qui sont comme devinées, induites.

### • La mémoire cognitive

C'est la première sorte de mémoire profonde, une sorte de zone de stockage de mots (au sens de la théorie du langage c'est-à-dire le couple signifiant/signifié, on voit un mot et on sait ce qu'il veut dire) ; la mémoire de travail fait constamment appel à la mémoire profonde pour identifier les mots connus.

### • La mémoire affective

C'est la deuxième sorte de mémoire profonde, plus complexe ; elle est couplée à la mémoire cognitive et fonctionne à un niveau inconscient, elle produit des images mentales qui rattachent le mot, le couple signifiant/signifié, à l'histoire personnelle du lecteur.

## b) Le parcours de l'œil

L'œil n'épelle pas, il parcourt un article par saccades : il divise inconsciemment la ligne en groupe d'environ 10 lettres et passe d'un groupe à l'autre. C'est donc un mouvement de type : fixation, déplacement, fixation etc., jusqu'à la fin de la ligne et le retour à la ligne suivante. La fixation dure  $\frac{1}{4}$  de seconde tandis que le déplacement se fait en  $\frac{1}{40}$  de seconde.

Globalement l'unité de lecture est donc le mot : l'acte de lecture est un décodage intellectuel de signes perçus par l'œil pour en dégager la compréhension des mots et du texte.

On peut justifier cette thèse avec les travaux de Noam Chomsky<sup>1</sup> pour qui comprendre une phrase revient à en reconstruire la structure syntaxique : plus la phrase est longue, donc compliquée, plus il sera difficile d'en reconstruire l'arborescence grammaticale. C'est une théorie qui n'a jamais été vraiment prouvée mais elle est pleine de bon sens !

### • Le nombre magique de 7, plus ou moins deux

La capacité de mémorisation est évaluée à 7 éléments (lettre ou chiffre) en moyenne, avec une correction possible de plus ou moins deux. C'est ce que les linguistes comme G. A. Miller<sup>2</sup> appellent l'empan visuel (ou mnésique) : l'empan au départ c'est le mot qui désigne

<sup>1</sup> CHOMSKY Noam, *Réflexions sur le langage*, Flammarion Poche, 1997; 283 pages, prix Fnac 6,84 € ; Le Langage et la Pensée, Payot Poche, 1990; 144 pages.

<sup>2</sup> MILLER G.A., *The magical number seven*, Psychological Review 63, 81-94 ; cité in

la largeur de votre main ouverte, entre les extrémités du petit doigt et du pouce puis ce mot a été récupéré par les linguistes avec la notion d'empan visuel, c'est-à-dire le nombre de caractères que l'œil peut lire en une seule fixation et qui est donc en moyenne de 7 caractères. Évidemment, cet empan varie selon le lecteur : un lecteur qui aurait un empan de 12 mots est censé retenir 85% d'une phrase de 14 mots, 70% d'une phrase de 17 mots et 50% d'une phrase de 24 mots.

Enfin, pour relativiser tous ces travaux, retenons que retenir et mémoriser, cela ne veut pas dire forcément comprendre !

- **Le processus d'agrégation ou « chunking »**

Il est logiquement difficile de mémoriser une suite de 10 items comme 0 1 4 4 2 5 3 2 3 0. Mais si on groupe ces items, ce qu'on appelle le processus d'agrégation (chunking en anglais) on diminue le nombre d'items et on retient plus facilement surtout quand on connaît le contexte, en l'occurrence un numéro de téléphone : 01 44 25 32 30 (c'est celui de mon bureau).

Le processus de mémorisation repose tout simplement sur le groupement de l'information et pour une quantité d'informations de base égale, le nombre de ces paquets (chunks) influe directement sur la capacité de rétention.

Une technique américaine moderne d'écriture, le principe de la pyramide de **Barbara Minto**<sup>3</sup>, s'appuie sur ce processus de regroupement par niveaux. Il est frappant de constater que ces règles de base des processus de mémorisation ont été appliqués par BT pour ses numéros de téléphone qui sont groupés par 3 ou 4 numéros, après des études poussées menées par l'un des meilleurs spécialistes mondiaux des processus mnésiques<sup>4</sup>. Mais ils ont été apparemment ignorés par France Télécom...

Quoiqu'il en soit, l'hypothèse de l'influence de la longueur des phrases est en accord avec la théorie du discours de **W. Kintsch**<sup>5</sup>, largement répandue aujourd'hui, selon laquelle le lecteur commence par élaborer une microstructure en transposant le texte lu sous forme de micro-propositions. On va toujours dans le sens de la simplification.

Reste un problème de taille : si à chaque fois qu'on veut se livrer à une interprétation, on a besoin de l'Encyclopedia Universalis pour être sûr qu'on ne passe pas à côté d'une information utile à la compréhension du problème, cela risque de ne pas être très pratique. Il va donc falloir choisir, sélectionner les infos dont on a besoin : c'est ce qu'on appelle le principe de pertinence<sup>6</sup>.

### c) La loi de l'effort minimum

Il y a un texte de **Dan Sperber** et **Deirdre Wilson** qui l'explique très bien :

*« Qu'est-ce que la pertinence ? Le traitement de l'information par les êtres humains leur demande un certain effort mental et produit en eux un certain effet cognitif. L'effort demandé est un certain effort d'attention, de mémoire et de raisonnement. L'effet produit consiste en*

LABASSE Bertrand, *La lisibilité rédactionnelle*, Communication & langages n°121, 1999 ; [www.univ-lyon1.fr/spr/GEV/READIBILITY.html](http://www.univ-lyon1.fr/spr/GEV/READIBILITY.html) .

<sup>3</sup> MINTO Barbara, *The Pyramid Principle*, Prentice Hall Financial Times, London, 2002 (première édition en 1987)

<sup>4</sup> BADDELEY A. D., *Working memory*, New York: Oxford University Press, 1986; cité in LABASSE Bertrand, *Une dynamique de l'insignifiance*, Enssib, 2002.

<sup>5</sup> KINTSCH W., *Comprehension ; a paradigm for cognition*, Cambridge University Press, 1997 ; cité in LABASSE Bertrand, *La lisibilité rédactionnelle*, Communication & langages n°121, 1999 ; [www.univ-lyon1.fr/spr/GEV/READIBILITY.html](http://www.univ-lyon1.fr/spr/GEV/READIBILITY.html) .

<sup>6</sup> SPERBER Dan et WILSON Deirdre, *La pertinence : communication et cognition*, Minuit, 1989 ; cité in *Le langage*, Ed. Sc. Humaines, voir bibliographie.

*une certaine modification de croyances de l'individu : l'addition de nouvelles croyances, l'élimination de croyances antérieures ou simplement un affaiblissement ou un renforcement de croyances antérieures. »*

Ils en déduisent deux règles :

- plus l'effet cognitif produit par l'information sur le destinataire est grand, plus elle sera pertinente pour lui ;

- plus l'information lui demande un effort de traitement, moins elle sera pertinente.

Il faut s'appuyer sur ce principe sélectif qui est quasi-inconscient au niveau du langage pour en déduire une démarche plus organisée dans le traitement d'informations.

Retenez aussi de toute cette théorie que dans un échange il y a toujours une double intention: transmettre un message et convaincre son interlocuteur, ce qui n'est pas du tout la même chose. Cette double fonction existe toujours même dans un rapport qui se voudrait purement informatif: je vous informe d'un fait mais la façon dont je vous le dis présuppose votre compréhension, je vous attribue donc implicitement des connaissances, des croyances et même des états mentaux nécessaires à une bonne interprétation de mon énoncé.

La pertinence d'une information dépend de son contexte et ce contexte est en quelque sorte « fabriqué » par l'interprète, le destinataire, le lecteur, il est choisi par lui dans le sens qui lui donnera le maximum de compréhension c'est-à-dire le maximum d'information pour le moindre effort. D'où les possibilités de malentendu et de manipulation.

Enfin, tout ce processus de compréhension implique des modèles mentaux, qui sont en fait des représentations du monde, des analogies au monde, des représentations imagées. Le modèle mental, selon Philip N. Johnson-Laird<sup>7</sup>, c'est « *la façon naturelle dont l'esprit humain construit la réalité, en conçoit des alternatives et vérifie des hypothèses, lorsqu'il est engagé dans un processus de simulation de modèles.* »

C'est une autre différence fondamentale entre communication et langage : la communication est une transmission de modèles tandis que le langage est le moyen de cette transmission.

Pour les psychologues, la structure du modèle est similaire à l'état de chose représenté, avec des éléments et des relations. Dans la hiérarchie de ces modèles, on part de simples images mentales figuratives centrées sur le point de vue particulier de l'observateur jusqu'à des modèles abstraits avec des relations entre unités conceptuelles.

Enfin, le modèle peut aussi être hypothétique, construit à la demande, si on a vraiment envie de comprendre !

#### **d) Nous sommes tous des « deductive satisfacers »**

Dans la théorie moderne de l'information, avant qu'on parle de communication (au sens ou on parle d'un directeur de la communication), il y a déjà cette idée d'une adaptation au contexte et au destinataire. C'est l'introduction d'un humanisme qui n'existait pas auparavant parce que la théorie de base a servi à l'informatique: traduire des données en 0 et en 1, ça peut se faire sans sentiment. Aujourd'hui les croyances reviennent en force, ce qui peut parfois fausser le raisonnement ou le simplifier. On s'arrête dès qu'on croit avoir compris, on est tous un peu flemmard. C'est ce que dit en tout cas Philip N. Johnson-Laird : « *Les sujets qui parviennent à une conclusion putative qui coïncide avec leurs croyances auront tendance à arrêter de rechercher des modèles alternatifs qui pourraient réfuter leur conclusion, les gens sont des deductive satisfacers* ».

---

<sup>7</sup> JOHNSON-LAIRD Philip N., *La théorie des modèles mentaux*, in EHRlich M.-F., TARDIEU HH., CAVAZZA M. (sous la dir. de), *Les modèles mentaux, approche cognitive des représentations*, Masson, 1992.

### e) Des mots courts

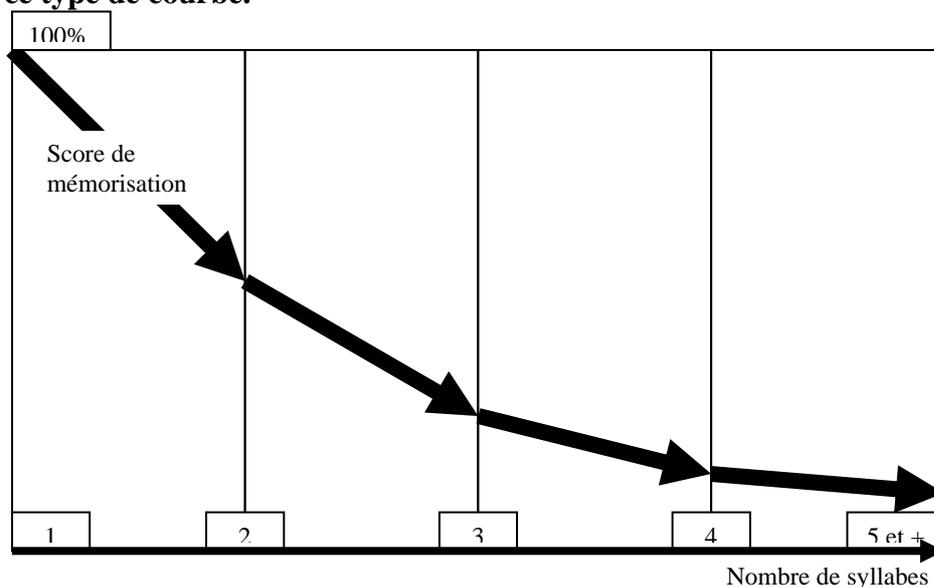
Donc, le lecteur lit et mémorise d'autant mieux que le mot est court et contient moins de syllabes. Plus le mot est long, plus l'œil sera obligé de faire un grand nombre de fixations. Ces fixations ne servent pas à identifier les caractères mais à permettre dès que possible l'accès lexical, c'est-à-dire qu'on va piocher dans notre dictionnaire, notre lexique mental et on va en ressortir le mot le plus plausible mais aussi le plus connu de nous: c'est pourquoi on aura tendance dans un premier temps à lire *électrique*, qu'on utilise souvent, plutôt que le mot écrit et qui est en fait *éclectique*, qu'on utilise moins. Si jamais la pioche est mauvaise, il faudra revenir en arrière (ce qu'on appelle la régression oculaire) et là le taux de lecture risque de fortement diminuer car c'est une manœuvre compliquée qu'on n'a pas envie de faire.

Ces enseignements ont notamment été confirmés par les travaux de deux spécialistes des signes, Pierce et Miller: ils nous expliquent que la lecture rapide est limitée par deux facteurs:

- **la capacité maximale du système de perception humain qu'ils évaluent, en reprenant les calculs de la théorie de Shannon à 45 bits par seconde<sup>8</sup>**; et c'est là qu'on se sent un peu faible par rapport aux machines: un disque dur d'ordinateur, a typiquement un taux de transfert de données de 30 millions de bits par seconde; on est nettement moins rapide, n'est-ce pas!

- **la capacité de la mémoire à court terme**: c'est la mémoire qui va servir la mémoire de travail mais qui travaille beaucoup plus vite, de l'ordre du centième de secondes. Certains auteurs estiment cette capacité **de 10 à 20 mots**, selon la qualité du lecteur.

**On pourrait symboliser la mémorisation des mots en fonction du nombre de syllabes par ce type de courbe.**



Vous pouvez vous amuser très facilement à calculer la longueur moyenne de vos mots dans un de vos textes, avec la fonction Outils/Statistiques de Word qui vous calcule le nombre de mots et de caractères (sans compter les espaces).

Par exemple, dans ce texte, Word compte 13 974 caractères (sans les espaces) et 4 927 mots, ce qui donne un mot d'une longueur moyenne de 4,87 caractères.

<sup>8</sup> cité in ESCARPIT Robert, *L'écrit et la communication*, PUF Coll. Que sais-je ?, 1984.

Quelques exemples célèbres de nombre de mots différents utilisés par de grands auteurs (exemples anciens tirés des manuels de journalisme, je n'ai pas eu le courage de les mettre à jour !)

- Charles De Gaulle : 6 009 (ensemble de ses discours)
- Françoise Giroud : 2 990 mots (sur 50 éditoriaux de L'Express)
- Jean Ferniot : 2 460 (23 chroniques à RTL)
- Jean Racine : 1 800 mots (ensemble de ses tragédies)
- Charles Trénet : 1 200 (ensemble des chansons)
- Georges Simenon : 800 (tous les Maigret)

#### • Nombre de mots compris par les lecteurs

On estime qu'un lecteur comprend et maîtrise en moyenne :

- 600 à 800 mots en fin d'études primaires (cible : lecteur du Parisien);
- 1 500 mots au niveau du bac ;
- 3 000 à 3 500 mots à un niveau d'études supérieures (cible : lecteur du Monde).

Par rapport à ces chiffres, on connaît et reconnaît 4 à 5 fois plus de mots mais nous n'en connaissons qu'un sens approximatif et nous ne les utilisons pas.

Rappelons que le Petit Robert contient 50 000 mots et que le thésaurus de la langue française en contient environ 150 000 mots, sans les jargons spécialisés, plus de 200 000 avec eux.

## f) La phrase

Les expériences de mémorisation de différents types de phrase sont sans équivoque. Si on tente de faire mémoriser deux phrases construites chacune en deux parties, mais l'une est simple et courte (12 mots : 6 + 6), l'autre très longue, avec des incises (rajout dans la phrase, sorte de parenthèse comme dans « *Un soir, t'en souvient-il, nous voguions en silence* » Lamartine, c'est ça que je viens de faire, une incise), le résultat est sans appel :

- pour la phrase courte, le taux de mémorisation de la 1<sup>ère</sup> partie de la phrase atteint les 100% et celui de la deuxième partie n'en est pas loin ;
- pour la deuxième phrase, on obtient les scores suivants ; 50% de mémorisation pour la 1<sup>ère</sup> partie et seulement 10% pour la seconde, soit 30% au total.

Cela nous donne le tableau suivant

**Tableau de mémorisation de la phrase en fonction de son nombre de mots**

Nombre de mots par phrase	Message entier retenu à...	Première moitié retenue à...	Deuxième moitié retenue à...
12	100%	100%	100%
13	90%	95%	85%
17	70%	90%	50%
24	50%	70%	30%
40	30%	50%	10%

On en déduit quelques règles simples :

- la mémorisation est une fonction décroissante quasi-linéaire du nombre de mots dans la phrase ;
- la première partie de la phrase est mieux retenue que la seconde ;
- les meilleurs scores de mémorisation se situent autour de 12 mots.

On comprend mieux en lisant ces chiffres pourquoi l'écriture journalistique privilégie les mots courts et les phrases courtes : c'est tout simplement une règle d'efficacité de lecture. Et pourquoi on enseigne que le sujet, le message essentiel de la phrase doit se situer au début : il a plus de chance d'être retenu et donc compris.

## g) Les critères de lisibilité

Il existe de nombreux travaux sur ce sujet, avec des tas d'indices plus ou moins scientifiques. Une analyse de **Bertrand Labasse**<sup>9</sup> : « En première analyse, ces indices reposent sur une règle de bon sens : le plus court est le mieux. On a, par exemple, ironisé sur le célèbre début de la "Recherche du temps perdu" ("Longtemps, je me suis couché de bonne heure...") qui annoncerait le caractère soporifique des phrases proustiennes. De fait, Proust est réputé difficile à lire : dans les premières pages de la "Recherche" figure une période de près de 500 mots. La longueur moyenne de ses phrases est de 38 mots contre 20 pour Gide et 13 pour les livres de la collection Harlequin (dixit Richaudeau, 1988, 1992). Il n'est donc pas surprenant que la formule de lisibilité de Flesch - celle qu'exploitent en général les programmes informatiques - crédite Proust d'un score... négatif (-10), contre 30 à Saint-Exupéry et 60 aux bandes dessinées Tintin et Spirou (De Landsheere, 1982). »

Voici, dans tous les travaux sur le sujet, les critères les plus cités<sup>10</sup> :

### • Au niveau du vocabulaire

- présence ou absence des mots utilisés dans une liste de référence, un vocabulaire de base variant de 500 à 7 000 mots suivant le contexte ;
- nombre de mots différents pour 1 000 mots consécutifs ;
- nombre de mots à plusieurs syllabes ;
- nombre de mots de plus de 3 syllabes ;
- nombre de lettres par mots ;
- nombre moyen de syllabes pour 100 mots ;
- nombre de mots d'une seule syllabe ;
- nombre de mots de plus de 6 lettres
- pourcentage de mots concrets ;
- nombre de mots dont la signification est modifiée par un adjectif ou un adverbe ;
- nombre de lettres, voyelles et consonnes ;
- nombre moyen de substantifs, d'adjectifs, de verbes auxiliaire.

### • Au niveau de la phrase

- nombre moyen de mots par phrase ;
- longueur des phrases les plus longues ;
- nombre de phrases à structure simple ;
- pouvoir séparateur des signes de ponctuation ;
- degré de prévisibilité des structures de la phrase.

### • Au niveau de l'intérêt humain

- nombre de mots sensoriels ;
- nombre de pronoms personnels ;
- nombre de mots personnels ;
- nombre de références personnelles ;
- nombre de phrases personnelles.

## h) Quelques formules de calcul de lisibilité

A partir de cette liste de paramètres, de nombreux chercheurs ont créé des formules pour mesurer le degré de lisibilité d'un texte. C'est bien sûr un exercice périlleux. Ces formules ne prennent pas en compte le sens réel du texte. Faites une liste de mots courts et de phrases

<sup>9</sup> LABASSE Bertrand, *La lisibilité rédactionnelle*, Communication & langages n°121, 1999 ; [www.univ-lyon1.fr/spr/GEV/READIBILITY.html](http://www.univ-lyon1.fr/spr/GEV/READIBILITY.html)

<sup>10</sup> Cité in HERVOUET Loïc, *Ecrire pour son lecteur*, ESJ Lille, Coll. J comme journalisme, 1979.

courtes sans aucun sens : le texte aura une très bonne note une fois passé dans ces différentes moulinettes.

Malgré ces réserves, cela peut être utile de les connaître.

- **Le « fox index » de Gunning**

Ou la cote de brouillard en français. Gunning est un consultant de presse américain ; sa formule = nombre moyen de mots par phrase + pourcentage de mots de trois syllabes et plus, le tout multiplié par 0,4 (un coefficient créé à l'époque pour que le nombre obtenu corresponde à une échelle éducative américaine).

Son échelle de valeur :

- 16 et + : Livres universitaires
- 12 et + : Commence à être difficile pour le grand public
- 10 : Score moyen des journaux américains
- 9 : Reader's Digest
- 6 : Bandes dessinées

En français, il faudra être moins sévère car les mots anglais sont plus courts et le taux de redondance de l'anglais est inférieur à celui du français. On peut prendre par exemple des mots de 4 syllabes et plus au lieu de 3.

Si j'applique la formule à une petite partie de mon texte de cours, j'obtiens une très bonne note de 8,3, je suis donc aussi clair que le Reader's Digest !

Certains chercheurs (comme Gilbert de Landsheere, directeur du laboratoire de pédagogie expérimentale de l'Université de Liège) conseillent de ne compter que les mots de 4 syllabes et plus en français, pour tenir compte de la différence avec l'anglais.

- **L'indice de Flesh (Reading Ease Level)**

Encore un consultant américain, Rudolph Flesh. C'était la formule la plus utilisée avant l'arrivée des ordinateurs et elle l'est restée depuis, intégrée dans de nombreux logiciels, comme Word de Microsoft.

Sa formule a l'avantage de la simplicité et elle peut être calculée par programme :

**206,835 – (1,015 x nombre moyen de mots par phrase) – (84,6 x nombre moyen de syllabes par mots)**

Le score varie de 0 (très difficile) à 100 (très facile).

- 100-90 Très facile
- 90-80 Facile
- 80-70 Assez facile
- 70-60 Moyen
- 60-50 Assez difficile
- 50-30 Assez difficile
- 30-0 Très difficile

Le Reader's Digest fait 65 et les bandes dessinées 90.

Mais c'est surtout une formule pour les textes anglophones.

Résultat : mon support de cours obtient 38, ce qui apparemment n'est pas très bon.

Aux Etats-unis, nous dit-on, ces indices de lisibilité sont devenus une contrainte légale dans les documents administratifs. Certaines formules ont été instituées en normes légales d'intelligibilité dans plus de la moitié des états américains : l'article 38a/699a de la législation du Connecticut stipule par exemple que les polices d'assurance doivent obtenir un score minimal de 45 au test de Flesh. Elles sont également utilisées dans les hôpitaux, pour contrôler la clarté des formulaires de « consentement éclairé » que doivent signer les patients avant une opération. (information tirée de Bertrand Labasse, non vérifiée).

• **Pourtant, il y a quatre obstacles :**

- toutes ces formules ne sont qu'approximatives sur le strict plan scientifique de la linguistique et de la phonologie ; le calcul de mots de Word se fait par exemple selon un chercheur<sup>11</sup> avec une approximation de 10%, ce qui est énorme ;
- elles ne sont qu'un comptage statistique et ne peuvent en aucun cas témoigner de l'intérêt véritable du texte ;
- la formule de Flesh peut se traduire par des nombres à plusieurs décimales : on ne voit pas très bien l'utilité pratique d'une telle précision, ce qui aurait tendance à invalider son caractère scientifique ;
- elles auraient tendance à mesurer plutôt le conformisme et la redondance puisque plus un texte est redondant, sans originalité, sans construction particulière, plus il serait facile à lire, alors que, dans la réalité, on risque vite de s'en désintéresser.

**i) Au fait, comprendre un texte, qu'est-ce que c'est ?**

Au-delà des formules diverses de comptage et de lisibilité, la vraie question est de savoir si l'on peut mesurer la compréhension d'un texte. On sait que comprendre un texte, c'est se construire une représentation, élaborer un modèle mental de ce qui est décrit.

Illustration de Bertrand Labasse qui compare deux énoncés ayant exactement la même structure syntaxique, syllabique et le même nombre de lettres :

*Le vagabond prend les gâteaux posés sur la fenêtre.*

*Le lamaneur voit le vraquier mouillé dans la darse.*

Ces deux phrases auraient le même score aux indices de lisibilité.

Pourtant on comprend vite la première, on imagine la scène, on suppose que le vagabond a faim, que ce ne sont pas ses gâteaux et qu'il les vole pour les manger. Imaginez la différence de sens si l'on avait mis « pâtissier » à la place de « vagabond ». Bref, il y a plein d'informations absentes que l'on remplace par des hypothèses et le texte est lu et compris vite.

Alors que dans la deuxième phrase, on va buter sur des mots peu connus, il faut être docker ou marin pour la comprendre : lamaneur=pilote chargé de l'entrée-sortie des bateaux dans les ports ; vraquier=navire transportant des produits en vrac ; darse=bassin abrité du port.

Le problème c'est qu'un mot peut être connu, par exemple parce qu'il va être répété par les médias, sans pour autant qu'on sache exactement ce qu'il veut dire : on va nous parler à la télé de *anticyclone* mais tout le monde ne saura pas ce que c'est et il y aura sans doute des contresens. Enfin, n'oublions pas le contexte qui aide à comprendre: le mot *bout* peut dire *fin* mais aussi *cordage* si on parle marine (auquel il se prononce *boutte*).

Bertrand Labasse propose une arborescence pour définir ce qu'il appelle l'intelligibilité d'un texte, qui prendrait en compte un certain nombre de paramètres :

---

<sup>11</sup> LABASSE Bertrand, *La lisibilité rédactionnelle*, [www.univ-lyon1.fr/spr/GEV/READABILITY.html](http://www.univ-lyon1.fr/spr/GEV/READABILITY.html)

			perceptive	<i>lisibilité graphique (optique)</i>
			acquisitive	<i>lisibilité lexico-syntaxique</i>
	difficulté		logique	<i>cohérence, interprétabilité</i>
			gnosique	<i>représentabilité</i>
pertinence	intelligibilité		métatextuelle	<i>finalité, genre, circonstances</i>
	attrait			
Typologie schématique de quelques facteurs intervenant dans la réception d'un texte				
Source: LABASSE Bertrand				

C'est une grille qui n'est pas terminée mais elle a le mérite de lister toutes les gradations de comportements que l'on peut avoir devant un texte. On part des difficultés de bas niveau (lisibilité graphique) pour aller jusqu'aux concepts, jusqu'aux niveaux qui dépendent essentiellement des connaissances et de l'implication du lecteur et qui vont finalement construire le sens.

Le niveau le plus bas inclut le choix des polices et des tailles de caractères. Peuvent s'y ajouter des éléments de maquette qui faciliteront ou compliqueront la lecture.

Le second registre de difficulté, celui de l'acquisition du contenu par le lecteur est sans doute celui où s'applique le mieux les formules de calcul de lisibilités que nous avons vues tout à l'heure. Par exemple, il est admis que le score à l'indice de Flesh rend assez bien compte du temps de lecture mais pas de ce qui est réellement retenu du texte. Mais cela peut quand même être un argument pour retenir ou acquérir des lecteurs : ainsi, on l'a vu, un lecteur de quotidien ne lit pas tout, loin de là ; or, selon un calcul fait J. Douël, cité par Bertrand Labasse si le journal diminuait un tant soit peu sa longueur moyenne de mots, par exemple en la ramenant de 4,8 à 4 caractères, le lecteur pourrait parcourir 20% de texte en plus à temps égal, ce qui donne davantage de chances de le séduire !

### j) Et quelques mystères pour finir

La linguistique obéit peut-être à des lois qui nous dépassent et de nombreux chercheurs ont trouvé des choses très étonnantes dans les textes : le linguiste américain Georges Kingsley Zipf a disséqué le vocabulaire de James Joyce dans « Ulysse » : pour ceux qui l'ont lu, ce n'est pas une mince affaire ! Je dirai que Joyce est le mélange détonant d'un Shakespeare moderne (il crée des mots tout comme lui, par exemple le mot « quark » qui a été choisi ensuite par les astrophysiciens pour désigner une particule élémentaire), de Proust pour la finesse des sens exprimés et de Céline pour la violence et la liberté de ton. Donc Zipf a trouvé que le 10<sup>e</sup> mot de Joyce dans Ulysse y revenait 2653 fois, le 100<sup>e</sup> 256 fois, le 1 000<sup>e</sup> 25 fois et le 10 000<sup>e</sup> 2 fois<sup>12</sup>.

Ce serait carrément une loi : la constance du produit de la fréquence d'utilisation du mot par son rang dans le texte. La loi serait valable également dans la langue française, selon Pierre Guiraud, cité par Loïc Hervouet<sup>13</sup>. Zipf explique cela en imaginant que l'auteur quel qu'il soit balance en permanence quand il écrit entre la loi du moindre effort poussant à employer les

<sup>12</sup> Cité in HERVOUET Loïc, *Ecrire pour son lecteur*, ESJ Lille, Coll. J comme journalisme, 1979.

<sup>13</sup> Cité in HERVOUET Loïc, *Ecrire pour son lecteur*, ESJ Lille, Coll. J comme journalisme, 1979.

mots les plus usuels et l'aspiration à l'originalité le poussant à chercher les mots les plus précis possibles.

Plus pratiquement, la formule de Zipf s'applique en fait à tout corpus de texte, pourvu qu'il soit un peu long et la règle est simple : si on classe les mots d'un texte en fonction de leur fréquence d'utilisation, on constate que le 100<sup>e</sup> mot de la liste sera employé environ 100 fois moins souvent que le premier, le 1000<sup>e</sup> mot sera employé 1000 fois moins que le premier, etc. Autrement dit le produit de la fréquence du mot par son rang est constant.

Selon l'étude menée par le professeur Jean Baudot<sup>14</sup> en 1992 sur la fréquence des mots de la langue française, le tiers d'un texte n'est composé que de 15 mots différents, ou encore, la moitié que de 47! C'est dire encore toute l'importance de la couverture lexicale des mots courants. Mais, inversement, dans un long texte, il y a beaucoup de mots à faible occurrence : toujours selon Jean Baudot, les trois quarts des mots du vocabulaire d'un texte d'environ un million de mots n'ont pas plus de 12 occurrences dans le texte et ne représentent que 5% des occurrences du texte<sup>15</sup>.

C'est le corpus mis au point par Baudot qui confirme notamment que les mots les plus utilisés dans la langue (et donc les mieux compris) sont les mots les plus courts : les 34 mots les plus utilisés ont une longueur moyenne de 2,8 lettres tandis que ceux situés autour de la 12 000<sup>e</sup> position comptent en moyenne plus de 8 caractères.

Labasse lui suggère de s'intéresser au ratio : nombre de mots différents/nombre total de mots, ce qu'il appelle le taux de redondance lexicale, ou « type token ratio ». L'idée est qu'un mot est plus facilement activé (lu, compris) si ce mot a déjà été rencontré dans le texte. Les textes pour enfants n'hésiteront pas ainsi à répéter quasiment des groupes entiers de mots d'une phrase à une autre.

Fin du cours sur la lecture ! La suite, l'écriture, c'est pour bientôt...

---

<sup>14</sup> BAUDOT Jean, *Fréquences d'utilisation des mots en français écrit contemporain*, Montréal, P.U.M., 1992.

<sup>15</sup> [www.osil.ch/eval/node29.html](http://www.osil.ch/eval/node29.html)