Du Texte à l'Image
From Text to Image

H. Marteau[1] A. Lefevre[1] N. Vincent[2]

[1]Laboratoire d'Informatique de l'Université de Tours
64 av Jean Portalis, 37200 Tours
hubert.marteau@etu.univ-tours.fr

[2]Laboratoire CRIP5-SIP, Université de Paris 5
45 rue des Saints Pères 75270 Paris Cedex 06
nicole.vincent@math-info.univ-paris5.fr

Résumé
Les problèmes créés par le nombre croissant de grands corpus textuels nécessitent une automatisation des traitements de comparaison de textes aussi variés soient-ils. Nous abordons le problème de manière statistique mais en conservant la possibilité de quantifier la structuration des textes. La méthode proposée repose sur une technique de transformation d'un texte quelconque en une image de taille fixe traduisant les éléments statistiques du texte et mettant en évidence des structurations étudiées à l'aide des outils de la géométrie fractale. Nous présentons une application à l'étude d'un corpus de 90 textes longs qui sont comparés et pour lesquels des conclusions quant à leur contenu peuvent être déduites de l'étude statistique.

Mots Clef
Classification de textes, indexation, fractale, changement de représentation.

Abstract
Problems created by the increasing number of large textual corpus make it necessary the automation of texts comparison treatments. We tackle the problem with a statistical way nevertheless we enable to preserve the possibility of quantifying the structuring of the texts. The proposed method relies on a technique which transforms a text into an image of fixed size. This image expresses the statistical elements of the text and highlights structures by the methods associated with fractal geometry. We present an application to the study of a corpus of 90 long texts which are compared together and for which conclusions on their contents can be deduced from the statistical study.

Keywords
Text classification, indexing, fractal, change of representation.

1 Introduction

Compte tenu des tailles des corpus textuels qui existent à l'heure actuelle et compte tenu de leur nombre en croissance perpétuelle, les outils d'indexation et de classification de textes sont de plus en plus nécessaires.
L'indexation a pour but d'associer aux entités un ensemble de paramètres, afin de permettre à tout utilisateur de retrouver facilement un document sur un sujet spécifique, c'est la recherche par mots clés ou de trouver à partir d' un texte donné un ensemble de textes ayant un certain lien avec le texte présenté, c'est la recherche par le contenu. L'indexation trouve de nombreuses utilisations que ce soit pour la classification de textes, la synthèse automatique, …
Si l'indexation, en ce sens, est une problématique essentielle, ce n'est pas la seule posée par la masse grandissante de textes à traiter.
Notre travail, s'il pourrait être utilisé en indexation de textes, s'inscrit dans le domaine de l'étude des textes et plus précisément de l'étude du lien qui peut exister entre le texte écrit et son contenu. Précisons qu'il s'agit de textes longs. Il s'agit donc pour nous de classer les textes selon des critères non nécessairement connus a priori.

1.1 Travaux antérieurs

On peut distinguer trois types d'indexation : l'indexation manuelle, l'indexation semi-automatique (supervisée) et l'indexation automatique (non supervisée) [5].
L'indexation manuelle nécessite qu'une personne désigne les termes d'indexation qui peuvent être des mots, des lemmes, des concepts, des n-grammes de caractères, … Il est évident que ce type d'indexation est de plus en plus difficile d'usage, il peut être appliqué sur de petits corpus limités, mais n'est pas imaginable pour un passage à l'échelle.
L'indexation semi-automatique propose à un utilisateur les termes d'indexation possibles selon leur fréquence, l'utilisateur n'a plus qu'à les accepter ou non.
Enfin l'indexation automatique ne demande aucune participation de l'utilisateur.

C'est évidemment ce dernier type d'indexation qui doit être mis en œuvre pour le traitement de gros corpus. Nous proposons néanmoins, pour certaines applications, de faire reposer cette indexation automatique sur une phase d'apprentissage. Un des éléments, essentiel et dual de l'indexation, est la recherche qui repose sur la comparaison entre les représentations choisies pour l'indexation.
Il existe de nombreuses méthodes de mesure dans l'espace des unités textuelles et des mots [3], nous pouvons en particulier citer le cosinus de Slaton [6], la distance du khi-deux, et le cosinus dans l'espace distributionnel.
Les méthodes d'indexation de textes se ramènent à un condensé d'informations des textes originaux : vecteurs, statistiques, par exemple. Ce condensé d'information sert de nouvelle représentation pour les textes de l'étude. Le problème de cette nouvelle représentation est qu'elle n'admet aucune limite de taille prévisible. En effet, on ne peut dire à l'avance le nombre de termes d'un texte, on ne peut pas déterminer à l'avance le nombre de liaisons entre les termes présents dans le texte, … On peut peut-être se limiter à un nombre d'informations, mais pour la comparaison de deux textes, l'un court et l'autre bien plus long, cette troncature conservera l'ensemble des informations pour le texte court et une quantité trop faible d'information pour le texte long. Les changements de représentations tels que les vecteurs, les statistiques, … permettent certes de condenser l'information, mais ne permettent pas d'obtenir une quantité fixe d'informations.

1.2 Contribution

Afin de remédier à ce problème, nous proposons de transformer un texte en image. C'est une idée originale car a priori un texte apparaît comme un signal monodimensionnel alors qu'une image est plane. Les voisinages usuels d'un point n'ont pas du tout la même structure dans les deux types d'espace mono et bi – dimensionnels. Nous avons choisi d'associer une image de taille fixe à tout texte. Ceci assure une normalisation des textes et rend plus aisés les traitements et les comparaisons qui seront alors réalisés dans un espace de dimension finie. D'autre part, après transformation du texte en image, nous pouvons appliquer les nombreuses techniques développées dans le domaine du traitement d'images.
Cette méthode a donc l'avantage de pouvoir être appliquée aussi bien sur des textes assez longs comme des œuvres littéraires que sur des textes plus courts comme un entretien à questions ouvertes ou un discours politique. Cette méthode doit même pouvoir être appliquée à des textes « abstraits » comme un texte représentant un son [1] ou un texte représentant une image [4].
Dans la partie qui suit nous présentons, après avoir défini l'unité textuelle choisie, le prétraitement que nous effectuons sur le texte, c'est à dire la technique de transformation du texte en image que nous proposons.
Dans la partie 3, nous présentons la méthode d'indexation que nous proposons. Enfin dans la quatrième partie, nous détaillerons un peu plus une application et le corpus associé et nous détaillerons les résultats que nous avons obtenus.

2. Du texte à l'image

Le texte, dans sa structure même, comporte un aspect temporel très marqué, et il est nécessaire de lire le texte entier pour se forger une impression globale. Au contraire, l'observation d'une image laisse instantanément une sensation globale. Evidemment les études occulométriques montrent qu'un balayage temporel contextuel a lieu mais sa rapidité n'en laisse aucune conscience à l'observateur. Nous cherchons donc, par une transformation, à créer une image à partir d'un texte qui puisse mettre en évidence certaines structures.
Nous présentons dans cette partie une méthode qui permet de transformer un texte en image. Cette méthode cherche à s'appliquer sur tout type de texte et peut donner plusieurs images en fonction des prétraitements effectués.
Une image peut être considérée comme un ensemble de motifs placés de manière harmonieuse dans un univers à deux dimensions. Nous présenterons donc dans un premier temps quelles sont les unités textuelles représentées par les motifs de l'image, puis nous expliquerons les raisons et la méthode du placement harmonieux des motifs assez simples, enfin nous verrons comment reproduire le traitement avec des motifs beaucoup plus complexes.

2.1. Les unités textuelles choisies : les n-grammes

La construction de l'image se fait pixel par pixel et on peut dire que chaque pixel représente un motif textuel. Chacun des motifs représente une ou plusieurs informations du texte original qui est représenté par l'image. Nous définissons donc par unité textuelle les éléments de texte qui seront représentés sur l'image. C'est à dire le type de découpage à effectuer sur le texte afin d'obtenir une collection d'objets ayant les mêmes propriétés. Il existe de nombreuses unités textuelles possibles, parmi les plus usuelles on peut compter les phrases, les mots, les lemmes, les concepts, les n-grammes, les lettres.
Nos réactions d'humain face à un texte nous laissent à penser que la langue, et par là même le texte, est une suite de lettres et de caractères, qui font sens par la présence d'une structure et nous interpréterions le texte à partir des mots constitués de caractères consécutifs. Nous avons choisi de prendre les n-grammes de caractères comme unités textuelles.
Un n-gramme est une suite de n caractères consécutifs.
L'ensemble des n-grammes, contenu dans un texte, peut ainsi être listé en faisant glisser une fenêtre de n cases sur le texte étudié.

Par exemple pour le texte « Je ne pense pas ! », on obtiendra les douze 6-grammes suivants : « je_ne_ », « e_ne_p », « _ne_pe », « ne_pen », « e_pens », « _pense », « pense_ », « ense_p », « nse_pa », « se_pas », « e_pas_ », « _pas_! » (les espaces ont été remplacés par des « _ » pour une plus grande lisibilité).
Il devient clair que ce mode de lecture du texte est une simplification puisque, ainsi, tout en gardant l'ensemble des informations, les entités ont toutes la même longueur. Il n'y a donc pas d'unités lexicales ayant une longueur imprévisible ou illimitée.
De plus, ce traitement permet un parcours très rapide du texte, il n'y a pas de découpage à effectuer (découpage en mots, découpage en phrases par exemple) comme pré-traitements.
Jalam et Chauchat [2] rappellent d'autres avantages que présentent les n-grammes. Tout d'abord on peut dire que les n-grammes de caractères ont le gros avantage de ne pas nécessiter de lemmatisation : on obtient directement la racine des mots sans prétraitement. En effet, si on lit les mots « pense » et « penser » en utilisant une fenêtre de 4 caractères, on voit que les deux mots ont en commun les 4-grammes « pens » et « ense », et le deuxième mot a le 4-gramme supplémentaire « nser », on arrive donc à trouver un fort taux de similarité entre ces deux mots juste en parcourant le texte sans prétraitement ce qui conserve la rapidité de lecture du texte. Le système permet donc aussi d'être tolérant aux fautes d'orthographe et aux déformations éventuelles dans une limite raisonnable.
Enfin, les n-grammes ne sont pas attachés à une langue particulière, ainsi la méthode peut être appliquée sans modification dans n'importe quelle langue utilisant un jeu de caractères ASCII standard. C'est à dire qu'elle ne s'appliquerait pas en l'état sur certaines écritures asiatiques.
Cette liste d'avantages n'est pas exhaustive, elle permet en particulier d'illustrer l'intérêt de la lecture par n-grammes par rapport à d'autres modes de parcours du texte.
Les n-grammes ont surtout l'avantage de considérer un texte comme on considère une partition de musique, c'est à dire qu'une lettre est indépendante, quand on considère l'ensemble du texte, des lettres qui la précèdent. Le texte est donc considéré comme un bloc unique, inséparable.
Ainsi chaque pixel est représentatif d'un n-gramme, sachant que chaque n-gramme est, par sa fréquence d'apparition, représentatif d'une quantité d'informations contenue dans le texte original. Cette quantité d'informations est mise en valeur par le niveau de gris du motif qui lui est associé. Le motif représentant le n-gramme le plus fréquent est donc de couleur noire, et le motif représentant les n-grammes absents sont de couleur blanche. Pour l'ensemble des autres motifs, une échelle de niveaux de gris est construite en fonction de la fréquence du n-gramme le plus fréquent. L'ensemble des motifs est donc issu du texte d'origine.

Afin de limiter le nombre de motifs possibles, l'alphabet, c'est à dire l'ensemble des caractères représentés lors de la lecture du texte original doit compter un nombre limité d'éléments.

2.2. Réduction de l'alphabet

Nous appelons alphabet le jeu complet de caractères de l'ensemble des textes étudiés par exemple l'alphabet français est constitué de 26 lettres en minuscule, 26 lettres en majuscule, de 10 caractères numériques, de lettres accentuées, de marques de ponctuation, …
Sachant que les alphabets suivent quelques variations avec les langues et avec le formatage des textes (certains ne sont écrits qu'en majuscules, certains sont écrits sans accents, …) nous avons préféré favoriser l'importance de la longueur des n-grammes plutôt que la quantité de caractères différents.
De plus, nous avons trouvé important de conserver à l'image le plus de symétries possibles, c'est pourquoi nous avons choisi de créer une image carrée. Cela nécessite que l'alphabet utilisé dans le texte ait une cardinalité ayant une racine carrée entière.
Compte tenu que l'alphabet utilisé en langue française (et anglaise) est composé à la base de 26 lettres (c'est à dire sans accent, sans majuscule) nous avons choisi de prendre un alphabet de 25 caractères. Cela nous permet de limiter le nombre de lettres tout en gardant les plus représentatives.
Pour cela, nous nous sommes, tout d'abord, limités à 24 lettres minuscules sans accents (les majuscules sont donc transformées en minuscules, et les lettres sont désaccentuées). De plus, nous avons groupé la lettre « y » avec la lettre « i » et la lettre « w » avec la lettre « v ». A ces 24 lettres, nous avons ajouté un caractère supplémentaire que nous appelons le caractère spécial (que nous représenterons par la suite par le caractère « & »), ce dernier caractère remplace tout ce qui, dans un texte, n'est pas caractère littéral, c'est à dire les espaces, la ponctuation, les chiffres, … Il y a ainsi 25 éléments de base possibles dans ce nouvel alphabet.
Cette réduction d'alphabet donne plus de poids à la lecture en n-grammes, en effet, le nouvel alphabet réduit le formatage du texte et permet à la méthode de lecture d'être plus tolérante aux erreurs sur ce point. Cette réduction a aussi pour avantage de fixer des limites prévisibles quant aux nombres de n-grammes possibles (25^n).
Chaque image est donc un ensemble de motifs dont le niveau de gris représente la fréquence d'apparition d'une suite de caractères recodés. Pour construire l'image il nous faut maintenant positionner les différents motifs dans un espace à deux dimensions. L'objectif est de mettre en évidence, grâce au principe de la localisation, les structures présents dans un texte.

2.3. Organisation de l'image : le patron

Il existe deux types d'univers d'application pour l'indexation : l'univers ouvert et les univers clos.
Un univers ouvert est un corpus dans lequel peuvent venir s'ajouter des textes, c'est le cas d'Internet par exemple, les moteurs de recherches qu'on y trouve ont donc une application dans un univers ouvert : pour leur indexation ils ne se basent pas seulement sur les pages qu'ils ont au départ, mais ont une architecture qui leur permet de faire face à l'arrivée de nouveaux textes qui peuvent ou bien ressembler à des textes déjà existants, ou bien être complètement différents.
Un univers clos est un corpus qui n'est pas destiné à s'agrandir. C'est le cas des études que nous menons. Nous cherchons à l'intérieur d'un corpus bien précis à comparer les textes entre eux, c'est à dire les images entre elles.
Cette notion d'univers d'application est très importante surtout dans le cas d'un changement de représentation aussi radical que le nôtre. En effet, la méthode est une méthode graphique qui s'applique sur les images représentant les textes du corpus. Pour que la méthode soit cohérente il faut qu'elle soit appliquée sur des images ayant au départ la même structure sous-jacente, c'est à dire la même organisation interne. Pour que cela soit possible, il faut que le schéma d'organisation des motifs, le patron, soit le même pour toutes les images du corpus.
Or ce patron est très important car c'est lui qui donnera un sens local et global aux images créées. C'est ce patron qui fera ressortir de l'image un ensemble structuré représentatif du texte ou un ensemble désordonné de motifs.
Nous allons à présent expliquer comment est construit ce patron pour des motifs représentant des 1-grammes, c'est à dire qu'on considère que chaque texte du corpus est lu avec une fenêtre ne contenant qu'une seule case, il est épelé.
Notre but étant de créer une image carrée et l'alphabet étant constitué de 25 caractères différents, il suffit d'organiser les motifs représentant ces caractères dans un tableau de pixels de taille 5x5. L'organisation des motifs n'a aucune raison de suivre l'ordre alphabétique, il n'a jamais été dit que cet ordre était représentatif des textes, qu'ils soient en français, en anglais, ou même en latin. Afin que le patron soit représentatif du corpus sa création dépendra donc du contenu du corpus : c'est la fréquence d'apparition des caractères dans la totalité des textes de l'étude qui indique l'ordre à suivre dans le tableau. Le motif représentant le caractère le plus fréquent est ainsi placé dans la case en haut à gauche du tableau et celui représentant le caractère le moins fréquent dans la case en bas à droite. De plus, le motif placé en haut à droite représente un caractère plus fréquent que celui du motif placé en bas à gauche.
Plus précisément, nous indiquons sur la figure 1 l'ordre utilisé pour le placement des motifs du plus fréquent au moins fréquent.

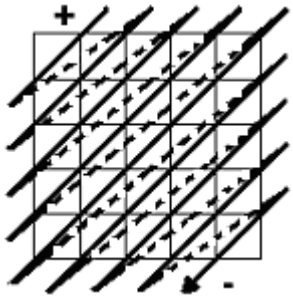

figure 1 : Organisation des motifs suivant leur fréquence d'apparition

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

figure 2 :Patron obtenu à partir des 1-grammes pour le texte : « Entretien n°1 »

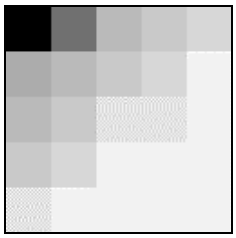

figure 3 : Image créée à partir des 1-grammes du texte : « Entretien n°1 » (taille réelle 5x5 pixels)

La figure 2 présente le patron obtenu sur notre corpus d'application, la figure 3 montre l'image obtenue pour un texte de l'étude, cette image ne montre pas un dégradé auquel on aurait pu s'attendre avec la création du patron, cela est dû au fait que le patron est construit sur les statistiques générales, c'est à dire à partir de l'ensemble des textes de l'univers, et le texte représenté ici montre ses particularités par rapport à la tendance générale.
Nous avons choisi, dans le cas d'un univers clos, que le patron soit représentatif du corpus entier, donc une fois celui-ci adopté c'est le même patron qui sera appliqué sur l'ensemble des textes du corpus. On obtient ainsi une image par texte, sachant que chaque image est construite en suivant une structure commune qui reflète une certaine logique dans la disposition des motifs par rapport à leur fréquence.
Cette étude des 1-grammes a été utilisée pour différencier les langues utilisées dans les textes du corpus. En effet, d'une langue à l'autre, la statistique d'apparition des caractères les uns par rapport aux autres n'est pas la même : en français E / ASITN / RULO / D / CMP / VQGFBH / JX / YZ ; en anglais E / TA / ONISRH / LDCU / PFMW / YBGV / KQXJZ.

Ainsi avec une limitation aux 1-grammes l'ensemble des textes français auraient la même représentation, l'ensemble des textes anglais auraient la même représentation et les représentations des textes français seraient différentes des représentations des textes anglais. Nous montrerons comment cette application peut apparaître comme le cas le plus simple et limité de l'approche plus générale que nous proposons en utilisant des n-grammes.
Comme nous venons de l'écrire, les statistiques d'apparition des 1-grammes sont différentes d'une langue à l'autre, elles sont aussi différentes suivant la nature des textes (littéraire, scientifique, …), or le but de notre méthode est d'être ouverte à un maximum de corpus, c'est-à-dire que nous souhaitons pouvoir appliquer notre méthode indépendamment de la langue, nous espérons même appliquer notre méthode sur des textes non littéraires [1], nous ne souhaitons donc pas fixer le patron a priori.

2.4. La répétition du patron

Nous avons vu précédemment la manière d'organiser les motifs représentants des 1-grammes de caractères pour former le patron de l'image. Pour le passage aux n-grammes de caractères, la méthode est identique.
Puisque l'ordre des fréquences d'apparition des 1-grammes de caractères est considéré comme un ordre naturel pour le corpus, c'est cet ordre qui déterminera aussi le patron des n-grammes de caractères (pour tout n). C'est donc le même patron qui est appliqué de manière récursive à l'intérieur de ses propres cases. On s'approche ainsi de la construction d'une image fractale. Nous présentons en figure 4, la répétition du patron à une échelle d'observation plus précise.

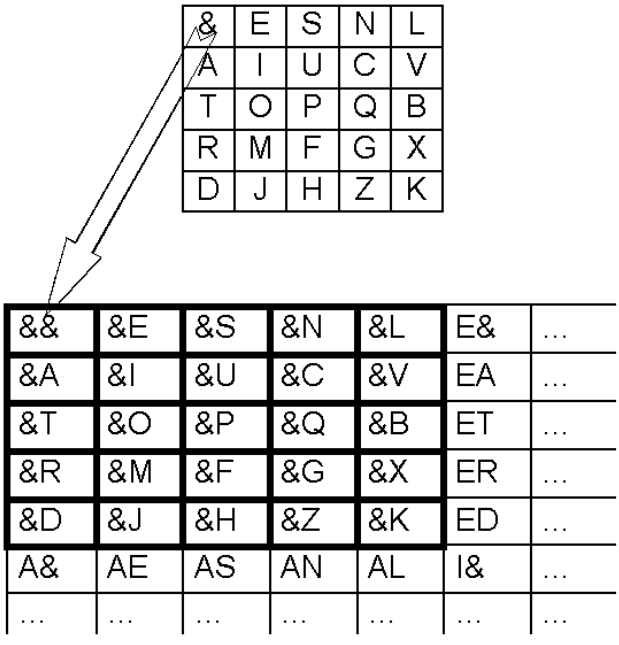

figure 4 : Récursivité du patron

Ainsi pour trouver l'information relative au motif « ab », il faut tout d'abord aller dans la case correspondant au caractère « a » (1ère colonne, 2ème ligne), puis dans cette case, il faut aller dans la case correspondant au caractère « b » (5ème colonne, 3ème ligne). Les coordonnées du motif « ab » dans le patron des 2-grammes de caractères (dimension : 25x25) est donc 5ème colonne, 8ème ligne. Nous présentons avec les figures suivantes les images d'un même texte obtenues pour différentes longueurs des n-grammes de caractères.

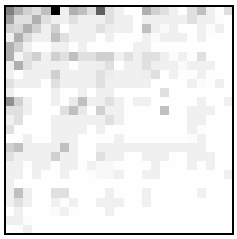
figure 5 : Image créée à partir des 2-grammes du texte : « Entretien n°1 » (taille réelle 25x25 pixels)

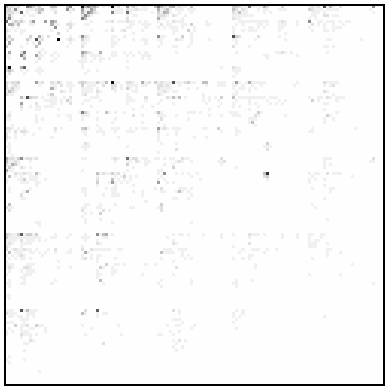
figure 6 : Image créée à partir des 3-grammes du texte : « Entretien n°1 » (taille réelle 125x125 pixels)

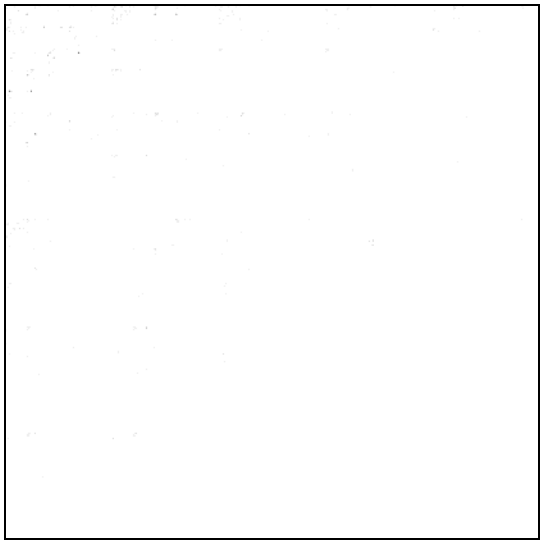
figure 7 : Image créée à partir des 4-grammes du texte : « Entretien n°1 » (taille réelle 625x625 pixels)

La figure 7 paraît pratiquement entièrement blanche, c'est une conséquence normale de l'évolution de la longueur des n-grammes. En effet, avec l'accroissement de la longueur des n-grammes, c'est-à-dire l'augmentation de n, le nombre de motifs possibles devient de plus en plus important : 25^n. Ainsi pour n = 4 il y a 390625 motifs possibles, alors que seuls 4397 motifs différents sont présents dans le texte : si on regroupait l'ensemble des pixels de l'image dans une zone précise de l'image, cela ne couvrirait que 1/ 89 de l'image, c'est à dire juste un peu plus de 1%.
De plus, une autre raison réside dans la distribution des fréquences d'apparition. Il y a en effet une augmentation du rapport entre la fréquence du n-gramme le plus fréquent et celle des autres. Par exemple le 4-grammes le plus présent dans le texte « Entretien n°1 » apparaît 125 fois (c'est le 4-grammes « que& »), soit un niveau de gris de 0 (c'est-à-dire noir). Le second plus fréquent n'apparaît que 116 fois (c'est le 4-grammes « ais& »), soit un niveau de gris de 18.
Sur les 4397 motifs différents, celui d'ordre 50 est « &une » et sa fréquence absolue est de 36 correspondant à un niveau de gris de 181 que nous percevons blanc. La figure 8 montre plus précisément l'évolution de la fréquence d'apparition des 4-grammes selon leur rang.

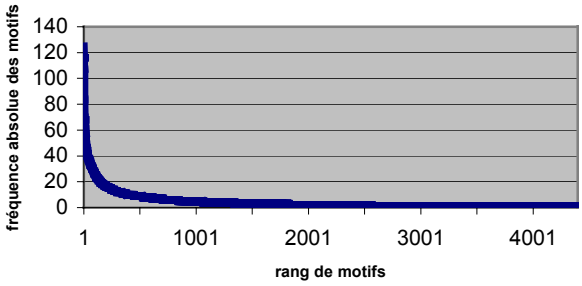

figure 8 : Evolution de la fréquence d'apparition des 4-grammes par rang

Le rapport entre le nombre de motifs différents et le nombre de motifs possibles ne pouvant être changé (on ne peut pas augmenter le nombre de motifs présents dans le texte), on peut tout de même changer le mode d'association des niveaux de gris attribués aux motifs. Une transformation linéaire conduit à la figure 7, nous avons rééchelonné l'intensité des motifs avec une double fonction logarithmique. La figure 9 montre l'image créée après le rééchelonnage et doit être comparée à la figure 7.

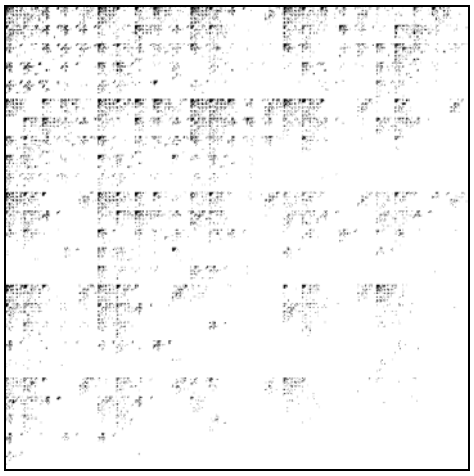

figure 9 : Image créée à partir des 4-grammes en utilisant un double rééchelonnage logarithmique du texte : « Entretien n°1 » (taille réelle 625x625 pixels)

Cette dernière image représente de manière explicite une grande quantité d'informations. Elle fait apparaître sa construction fractale. De la même manière, une image peut être créée pour tout texte, les images étant créées, il ne reste plus qu'à les comparer de manière à réaliser une classification des images entre elles, c'est-à-dire des textes entre eux.

3. La méthode d'indexation

L'image est construite de manière récursive par répétition du même patron, cela nous conduit à une image de type fractal. Nous avons donc choisi d'appliquer une méthode basée sur la dimension fractale de l'image comme méthode d'indexation.

3.1. La dimension fractale

La dimension fractale d'une image mesure la façon dont la fractale occupe l'espace de l'image, la dimension sera donc unique pour chaque image.
Pour un segment de droite, ensemble géométrique autosimilaire, on peut le considérer comme l'union de 3 sous-segments identiques entre eux et semblable au segment initial, la dimension fractale est de 1 (longueur L = 1 du segment, nombre N = 3 de sous-segments, log(N)/log(N/L) = log(3)/log(3/1) = log(3)/log(3) = 1) ; pour un carré constitué de 9 sous-carrés identiques, la dimension fractale est de 2 (taille L = 1 du carré, nombre N = 9 de sous-carrés log(N)/log(N/L) = log(9)/log(9/3) = log(9)/log(3) = 2).
Dans le cas d'ensembles plus complexes la notion de dimension d'autosimilarité se généralise par exemple en dimension de masse.

A partir d'un point d'un ensemble X et en notant $m(X_k)$ la mesure de $X_k = X \cap [\ 0\ ;\ k\ ]^2$ l'expression de la dimension de masse devient :

$$D = \lim_{k \to 0} \frac{\ln(\ m\ (X_{\ k}))}{\ln(\ k\ )} \quad (1)$$

3.2. Adaptation à notre problème

Comme il a été expliqué dans la partie précédente, la dimension fractale est une mesure d'occupation de l'espace. Cependant, dans notre méthode, il ne s'agit pas d'un segment constitué de sous-segments, mais d'une image qui semble constituée de sous-images et l'unité de base de chaque image est le point, ou pixel. La masse pour un tel ensemble se calcule donc à partir des points.
De plus, l'image est la représentation d'un texte et les points sont les représentations d'unités textuelles, la masse de l'image est donc une mesure d'utilisation d'un type d'unité textuelle et de sa structuration.
L'image apparaît visiblement de type fractal, on peut calculer la masse sur l'image entière. On peut aussi considérer les sous-images. La masse calculée sur une sous-image permet de mesurer la fréquence d'apparition d'un sous-ensemble de n-grammes. Par exemple pour une image $[\ 0\ ;\ 625\ ]^2$ construite à partir des 4-grammes, la sous-image $[\ 0\ ;\ 125\ ]^2$ rassemble les points correspondant aux motifs commençant par le 1-gramme le plus fréquent, la masse de cette sous-image renvoie donc la fréquence d'apparition de l'ensemble des 4-grammes compris dans ce sous-ensemble.

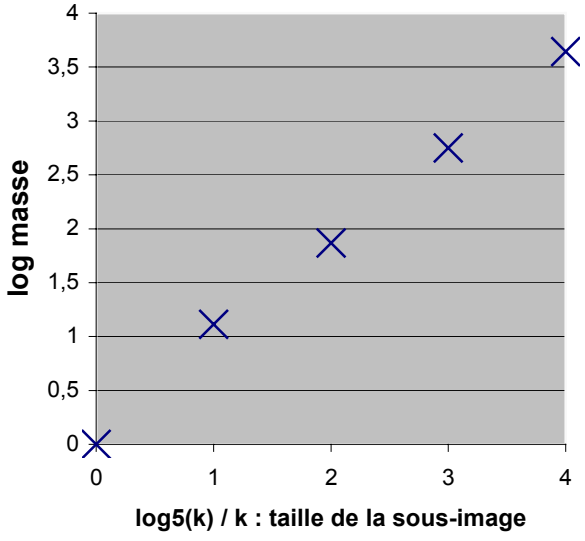

figure 10 : Augmentation de la masse sur différentes sous-images de l'image construite pour le texte « entretien n°1 »

La figure 10 montre l'évolution de la masse pour les sous-images représentant les sous-ensembles de 4-grammes commençant par « &&&& », « &&& », « && », « & », « », la dernière sous-image est l'image entière.
On constate que l'évolution que nous avons représentée sur la figure 10 est quasi-linéaire. Le coefficient de régression de la droite formée à partir de l'évolution des masses exprime la complexité de l'image. Si ce coefficient de régression avait pour valeur 0, cela indiquerait qu'il n'y a qu'un unique point dans l'image et que ce point est situé en haut à gauche de l'image. De même, si le coefficient de régression avait une valeur de 2, cela signifierait que l'image est entièrement noire en tout point. Le coefficient de régression de la droite formée par l'évolution des masses, ou dimension de masse, correspond donc à une valeur d'autosimilarité représentative de l'image. C'est cette linéarité que nous observons qui rend possible le calcul de la limite qui figure dans la formule (1).
La dimension de masse étant calculée à partir d'une droite, elle ne nécessite donc que trois points pour être calculée. Il est possible de considérer l'image globalement mais aussi des sous-images et de calculer pour chacune une dimension de masse. Nous avons fait plusieurs calculs. Les dimensions de masse calculées étaient celles de l'image entière, et les 25 dimensions de masse des sous-images directes de l'image, c'est-à-dire, pour chaque caractère, la dimension de masse qui correspond à l'ensemble des n-grammes commençant par le caractère établi. Il s'est avéré que la plupart des dimensions de masses des sous-images étaient proches de la dimension de masse de l'image globale. Les valeurs non représentées par la valeur de l'image globale sont celles qui correspondent à des sous-images représentant des n-grammes dont la première lettre est peu fréquente comme le K et le Z ou dont l'utilisation est bien particulière comme le Q qui s'utilise généralement suivi d'un U. Les résultats étant les mêmes pour l'ensemble des images du corpus, nous avons considéré comme valeur d'indexation la dimension de masse associée à l'image globale.

3.3. Longueur

La méthode est basée sur les n-grammes, il est donc judicieux de se demander quelle longueur n utiliser. Il est évident que l'on souhaite la plus grande discrimination possible afin d'associer le plus grand écart possible entre deux textes fortement dissemblables. Nous avons étudié, sur un corpus varié quelconque, l'amplitude entre les deux textes extremums en fonction de la longueur n choisie. La figure 11 montre le résultat obtenu.
Il apparaît que plus la longueur n grandit plus l'écart entre les valeurs associées aux images extrêmes est grand. Ce résultat était prévisible car en augmentant la longueur, on augmente en quelque sorte la précision de lecture, mais surtout le nombre de motifs différents.

Cette croissance est lente, de type logarithmique, contrairement à la taille des images. En effet, les images ont une aire dépendant de n : 25^n, leur croissance suit donc une évolution de type exponentiel : si on applique la méthode avec une lecture par des 4-grammes de caractères l'image obtenue a une taille de 625x625 qui peut être jugée comme acceptable alors que si on l'applique avec une lecture par des 6-grammes de caractères l'image résultante a une taille de 15625x15625, et les temps de calcul plus longs d'autant.

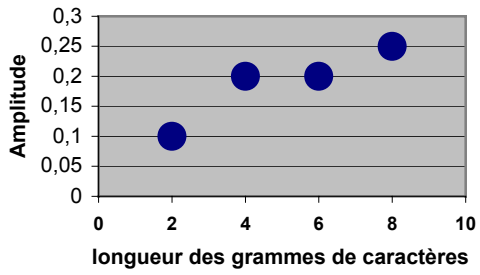

figure 11 : Evolution de l'amplitude de discrimination suivant la longueur des n-grammes de caractères

Ainsi compte tenu du rapport entre l'amplitude de discrimination et le coût qui comprend la taille de l'image (espace mémoire) et temps de calcul, nous avons fixé à 4 la longueur des n-grammes de caractères pour le type de discrimination que l'on souhaite réaliser. Nous verrons néanmoins, que pour une application plus simple comme la reconnaissance de langue, cette longueur peut être raccourcie.

4. Applications et Perspectives

Nous présentons dans cette partie deux applications, une première sur un corpus multilingue où nous avons juste créé les images à partir des 1-grammes de caractères, une seconde a été menée sur un corpus de textes sociologiques, puis nous commenterons notre choix de précision, enfin nous détaillerons nos perspectives.

4.1. Corpus multilingue

Le corpus d'application est un ensemble de quatre œuvres littéraires. Deux ont été écrites en français : « Notre Dame de Paris » de Victor Hugo et « Méditations Poétiques » de Alphonse de Lamartine ; les deux autres ont été écrites en anglais : « The Tragedy of Hamlet, Prince of Denmark » de William Shakespeare et « The Picture of Dorian Gray » d'Oscar Wilde.
Nous nous sommes limités à ce corpus parce que, nous le rappelons, l'identification de langue n'est pas l'objectif de la méthode, c'est juste le résultat de la représentation du texte.

Nous avons donc représenté les textes de ce corpus par des images construites à partir des 1-grammes de caractères. La figure 12 montre le patron obtenu pour cette construction.

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

figure n°12 : Patron pour la construction des images du corpus multilingue

Les figures 13 à 16 sont les différentes images obtenues.

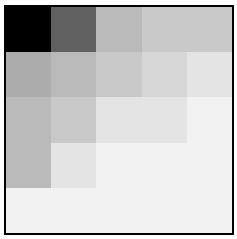

figure n°13 : Image créée à partir des 1-grammes de l'œuvre littéraire : « Notre Dame de Paris »

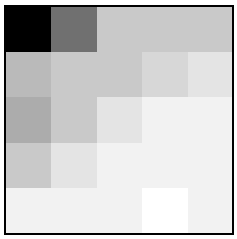

figure n°14 : Image créée à partir des 1-grammes de l'œuvre littéraire : « MéditationsPoétiques »

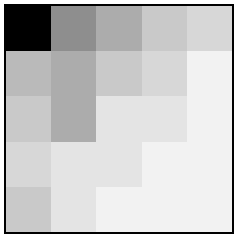

figure n°15 : Image créée à partir des 1-grammes de l'œuvre littéraire : « The Tragedy of Hamlet, Prince of Denmark »

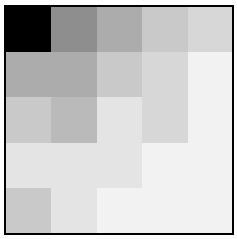

figure n°16 : Image créée à partir des 1-grammes de l'œuvre littéraire : « The Picture of Dorian Gray »

(taille réelle 5x5 pixels)

Cependant, pour ne garder que les différences les plus explicites, dans la première colonne en français les trois pixels centraux (ils correspondent en allant du haut vers le bas aux caractères : A,S,U) sont utilisés avec une fréquence presque similaire et le dernier (correspond au caractère : H) est très peu utilisé, en anglais le rapport diffère un peu, le troisième motif (U) est très peu représentatif alors que le dernier est bien représenté, ce qui est normal puisqu'il correspond au caractère H qui est utilisé en anglais par de nombreux mots : « the », « she », « they », « their », « who », … Des différences sont aussi visibles dans la deuxième colonne, mais ne seront pas détaillées ici.
On peut conclure pour cette courte étude qu'avec une faible précision (les 1-grammes de caractères) notre méthode modélise avant tout la langue utilisée.

Cette différence visuelle peut être quantifiée par un calcul de distance dans l'espace IN[25]. Dans l'étude qui suit, nous allons montrer l'application que nous avons faite avec une plus grande précision.

4.2. Application réelle

Le corpus d'application est constitué de 90 entretiens de type sociologique. Ces entretiens concernent 30 familles dans lesquelles ont été interrogés un enfant, la mère et le père. Ces entretiens avaient pour but de déterminer le passage des valeurs au sein d'une même famille.
Les thèmes abordés étaient la religion, l'éducation, l'autorité familiale, la politique et les travaux ménagers.
Nous avons appliqué la méthode avec une lecture des textes par 4-grammes de caractères et choix du patron.
La figure 17 indique le résultat obtenu par la méthode d'indexation pour chaque individu, les résultats sont organisés par famille et suivent, à l'intérieur d'une même famille, l'ordre énoncé précédemment : enfant, mère, père.
Le premier entretien remarquable est l'entretien n°33, cet entretien n'est pas un véritable entretien, il s'agit en effet d'une description résumée de l'entretien, et non d'une véritable discussion. Les phrases sont créées par l'interviewer et ne reflètent que le contenu de l'interview et en aucun cas la forme, le nombre de motifs est donc bien moins important que dans de véritables entretiens. La structure de l'image représentant l'entretien n°33 est donc bien moins complexe que la structure des images des autres entretiens.
A l'inverse l'entretien n°52 est l'entretien qui a la valeur la plus haute. Il s'agit d'un entretien fait avec un enfant qui suit des études de philosophie, son discours est assez complexe, avec un vocabulaire très varié, cette personne n'est attachée à aucune valeur en particulier et a plutôt un avis neutre sur l'ensemble des thèmes abordés. Les parents de cet individu, les entretiens n°53 et 54, ont un discours plus régulier, avec un vocabulaire peu varié, ils sont plus traditionalistes. Après analyse de l'ensemble des résultats on observe que les entretiens ayant une valeur assez élevée sont associés aux individus qui marquent une certaine indifférence quant aux sujets abordés, ils tiennent des discours assez complexes qui ont assez souvent tendance à s'éloigner de l'objectif principal des questions posées alors que les personnes attachées à certaines valeurs abordées dans les thèmes ont tendance à répondre plus directement aux questions.

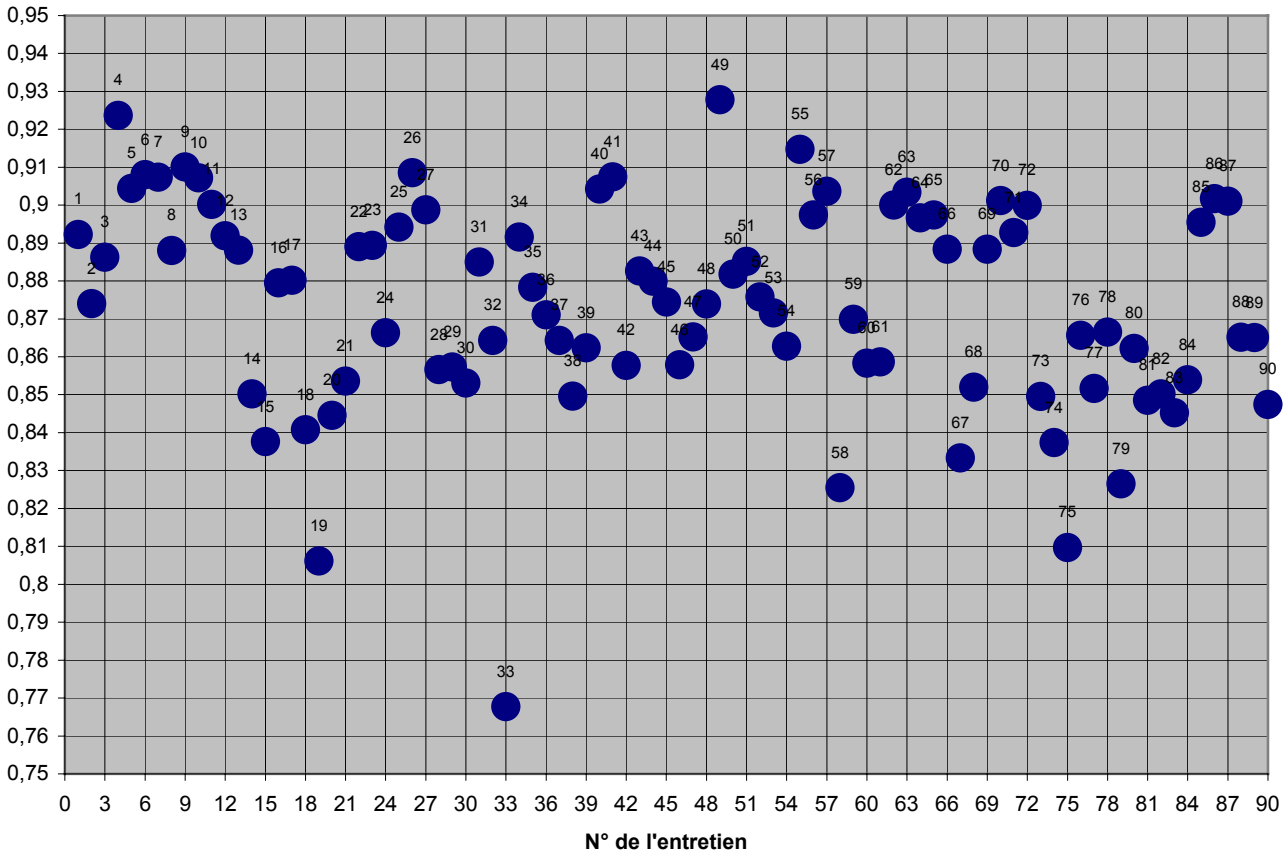

figure 17 : Dimension associée à chaque texte

Par exemple pour une question sur la peine de mort, dans la famille n°17, c'est-à-dire les individus n°52 à 54 vus précédemment, l'enfant répond à la question en 1343 caractères, la mère répond à la question en 481 caractères et le père répond à la question an 446 caractères.
Les parents ont un discours assez succinct et peu justifié, ils expriment juste leur opinion, l'enfant a plus besoin de se justifier et aborde différents cas où la peine de mort est compréhensible et les cas où elle ne l'est pas et même si il montre un attachement plus profond contre la peine de mort, son opinion est bien moins tranchée que celle de ses parents.

5. Conclusion

Nous proposons une nouvelle approche qui a pour but de changer la représentation de données textuelles en données graphiques. Ce changement de représentation a pour but l'application de méthodes et de mesures géométrique pour l'indexation des données originales.
Cette nouvelle représentation ainsi que la méthode d'indexation associée arrivent à distinguer les différentes structures et compositions des textes étudiés, ces deux méthodes peuvent donc être utilisées pour analyser des textes longs et en déduire des corrélations entre les caractéristiques du texte utilisé et des propriétés du contenu. Les textes d'études peuvent provenir de textes littéraires, entretiens avec questions ouvertes, textes abstraits ou codages, …
La méthode peut être appliquée à partir de nombreuses méthodes de représentation et on peut tester de très nombreuses méthodes de traitement d'images adaptées à la comparaison. Ainsi, cette méthodologie pourrait aboutir dans de nombreux problèmes liés aux textes comme l'identification d'auteurs ou la segmentation thématique par exemple.

Bibliographie

[1] E. Dellandrea, P. Makris, M. Boiron, N. Vincent, A medical acoustic signal analysis method based on Zipf law, *International Conference on Digital Signal Processing (DSP 2002), Santorini, Grèce,* Vol. 2. p. 615-618. Juillet 2002.
[2] R. Jalam, J. H. Chauchat, Pourquoi les n-grammes permettent de classer des textes ? Recherche de mots-clés pertinents à l'aide des n-grammes caractéristiques, *JADT 2002, St Malo France,* 13-15 Mars 2002.
[3] A. Lelu, Evaluation de trois mesures de similarité utilisées en sciences de l'information, *Les journées d'étude des systèmes d'information élaborée, Ile Rousse, France,* 15-18 Octobre 2002.
[4] N. Nikolaou, N. Papamarkos, Color image retrieval using a fractal signature extraction technique, *Engineering Applications of Artificial Intelligence*, Vol. 15, p. 81-96. 2002
[5] B. Pouliquen, Indexation de textes médicaux par extraction de concepts, et ses utilisations, *Thèse de doctorat à l'Université de Rennes 1*, 7 Juin 2002
[6] G. Salton, *Automatic Text Processing, Addison* Wesley, 1989.