L'automate textuel pour la prise en compte de l'évolution du texte

Hubert Marteau*, Nicole Vincent**

*Laboratoire d'Informatique, 64 av Jean Portalis, 37200 Tours
hubert.marteau@etu.univ-tours.fr
http://www.li.univ-tours.fr
**Laboratoire CRIP5-SIP, Université Paris 5, 45 rue des Saints Pères, 75270 Paris Cedex 06
nicole.vincent@math-info.univ-paris5.fr
http://www.math-info.univ-paris5.fr/crip5/

Résumé. Il n'est plus à rappeler que le corpus textuel, est tel qu'il est actuellement, intraitable à l'échelle et que sa croissance nous confirme l'obligation d'utiliser des outils automatiques de traitement. Cet article s'intéresse plus particulièrement à la caractérisation de textes et par là même à celle d'auteurs. A l'heure actuelle, toutes les méthodes existant travaillent sur un à-plat des textes traités. C'est-à-dire que ces méthodes travaillent sur le document fini, sans admettre qu'un cheminement existe entre le début du document et sa fin. Nous proposons une méthode tentant d'apporter cette notion d'évolution textuelle en traitant le texte par un automate. Cette méthode a pour but d'apporter des informations complémentaires quant au cheminement logique adopté lors de la création du texte. Nous présenterons les règles de l'automate et l'évaluation choisie. Puis nous présenterons des résultats validés par des experts, obtenus sur un corpus d'entretiens sociologiques.

1 Introduction

Le corpus textuel existant atteint une taille qui depuis longtemps le rend intraitable par l'homme de manière exhaustive. Sa perpétuelle croissance et le fait qu'aucune limite ne peut être envisagée à cette croissance confirment le besoin de traitement automatique de toute la quantité de données brutes présentes dans ce corpus. La première étape, comme pour tout problème de gestion de données, consiste à effectuer une classification des données.

La classification des données a pour but de créer un système dans lequel chaque donnée appartient à un ou plusieurs groupe(s) selon les informations que l'on souhaite traiter. (Mothe et al 2001) par exemple se limitent aux balises des textes traités, c'est-à-dire le titre, les auteurs, etc. … Il existe (Pouliquen 2002) trois types de classification : la classification manuelle, la classification semi-automatique et la classification automatique. La classification manuelle est le résultat d'un traitement par l'homme ; comme on l'a indiqué précédemment, elle est impossible à envisager. La classification 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. La classification automatique est réalisée de manière complètement automatique et a pour résultat, dans la plupart des cas, d'établir une distance entre les textes.

L'un des résultats les plus connus est sans doute celui de Salton (Salton 1971)(Salt 1989) et son vecteur de données (Singhal et Salton 1995).

D'autres méthodes utilisent une analyse de données basée sur les techniques de Benzécri (Benzécri 1973), Alceste (Reinhert 1995) est, dans ce domaine, certainement le logiciel le plus connu, il utilise une classification descendante hiérarchique qui trouve son application dans des domaines comme la sociologie, la psychologie, etc., pour le traitement de questions ouvertes (Garnier et Guérin-Pace 1998). D'une manière générale, l'analyse factorielle trouve son application tant pour le traitement d'œuvres littéraires (Ouazzani 1998) que pour le traitement thématique de textes ou le traitement de questionnaires à questions ouvertes (Morin 2002).

D'autres méthodes plus différentes existent. (SanJuan et Ibekwe-SanJuan 2002) proposent un système d'extraction de mots, de création de graphes, de recherche de composantes connexes et d'utilisation d'une classification ascendante hiérarchique. (Forest et Meunier 2002) proposent une méthode basée sur l'extraction d'unités lexicales et l'utilisation d'un réseau de neurones (ART1). (Bavaud et Xanthos 2002) proposent l'idée originale de se servir des lois de la thermodynamique.

Toutes ces méthodes sont basées sur l'extraction de composantes textuelles : mots, co-occurrences, unités lexicales, etc. De nombreuses méthodes préfèrent travailler sur les n-grammes. (Schbath et al 1995), (Schbat 1997), (Deschavanne et al 1999), (Dufraigne et al 2001), (Lespinats et al 2003), (Lespinats et al 2004) démontrent l'utilité des n-grammes en biologie pour la classification des espèces par leur chaîne d'ADN. De par leur possibilité de stockage (Gu et Berléant 2000) et leurs utilités en général (Jalam et Chauchat 2002), les n-grammes trouvent parfaitement leur place dans le textuel, que ce soit pour la recherche d'informations (Crowder et Nicholas 1997), en classification textuelle en général (Cavnar et Trenkle 1994), (Benhadid et al 1998), (Marteau et al 2003), (Marteau et al 2004) ou dans des cas plus particuliers comme la comparaison de requêtes (Karthik et Davis 2002) et la classification thématique, en particulier (Biskri et Meunier 2002).

Les méthodes liées à ces deux cas particuliers ont aussi beaucoup d'intérêt. Il existe des méthodes standard de similarité textuelle qui ont fait l'objet d'études comme celle de (Lelu 2002). Et en classification thématique on trouve une grande diversité de méthodes originales : (Ferret et al 1997), (Ferret et al 2001) qui ont une approche à la fois linguistique et statistique, (Boufaden et al 2001), qui repèrent des marqueurs oraux pour la segmentation de dialogues retranscrits, (Reynar 1994) qui a l'idée originale de représenter le texte sous forme graphique et le travail de (Hubert et al 2002) qui étudie l'apparition du nouveau vocabulaire. Cette dernière méthode ayant d'ailleurs fait l'objet d'une étude sur Corneille et Molière (Labbé 2003).

Nous proposons une méthode originale, ayant pour but d'effectuer une classification sur plusieurs niveaux d'entretiens oraux retranscrits. En effet, on ne s'intéresse pas seulement au contenu du texte mais aussi à sa construction. Pour cela, nous avons créé un automate qui prend le texte en entrée et crée une image représentant le texte d'un point de vue global.

Nous verrons dans la partie 2 comment est conçu l'automate. La partie 3 présente, quant à elle, les valeurs que nous extrayons au cours de la création de l'image et la méthode de classification que nous utilisons. Enfin, nous présenterons quelques résultats tendant à prouver la recevabilité de notre méthode.

Le domaine d'application de cette méthode est l'étude d'un ensemble de corpus d'entretiens de type sociologique.

Ces corpus sont indépendants les uns des autres puisque construits chacun pour une étude sociologique. Lors des entretiens, un ensemble de thèmes, fixé pour le corpus, est abordé.

2 L'automate de construction d'image

Avant de détailler notre méthode qui est basée sur un automate, nous allons faire quelques rappels sur les automates finis déterministes. Ensuite, nous détaillerons notre propre automate.

2.1 Rappels sur les automates

(Noureddine 1992) de manière intuitive, on peut voir un système de reconnaissance comme une machine permettant de lire un mot à travers différentes manipulations. Cette machine, appelée automate, permet donc, par extension, de reconnaître un langage.

Il existe plusieurs types d'automate, cependant ils ont tous une structure commune. Ainsi, un automate est composé de trois parties :

- Une bande en entrée, finie ou infinie, sur laquelle va s'inscrire le mot à lire. Une bande, en entrée, est divisée en cellules ; le mot à lire étant formé d'une suite de symboles (de l'alphabet), un symbole (et un seul) est logé dans une cellule.
- Un organe de commandes qui permet de gérer un ensemble fini d'états. La gestion des états se fait à travers une fonction spécifique, dite fonction de transition.
- Eventuellement, une mémoire auxiliaire de stockage.

(Aho et al 1986), (Noureddine 1992) Un automate fini non déterministe est un modèle mathématique défini par le cinq-uple $A(X,E,e_0,t,F)$, avec :

- X : un ensemble de symboles d'entrée (l'alphabet des symboles d'entrée)
- E : l'ensemble des états
- e_0 : un état qui est distingué comme l'état de départ ou état initial
- t : une fonction de transition, qui fait correspondre des couples état-symbole à des ensembles d'états
- F : un ensemble d'états distingués comme états d'acceptation ou états finals.

Un automate fini non déterministe peut être représenté graphiquement comme un graphe orienté étiqueté, appelé graphe de transition, dans lequel les nœuds sont les états et les arcs étiquetés représentent la fonction de transition. Ce graphe ressemble à un diagramme de transition, mais le même caractère peut étiqueter deux transitions ou plus en sortie d'un même nœud et les arcs peuvent être étiquetés par le symbole spécial ε (neutre) au même titre que les symboles d'entrée.

(Aho et al 1986), (Noureddine 1992) Un automate fini déterministe est un cas particulier d'automate fini non déterministe dans lequel :

- aucun état n'a de ε-transition, c'est-à-dire de transition sur l'entrée ε
- pour chaque état e et chaque symbole d'entrée a , il y a au plus un arc étiqueté a qui quitte e.

Un automate fini déterministe a au plus une transition à partir de chaque état sur n'importe quel symbole. Si on utilise une table de transition pour représenter la fonction de transition de l'automate fini déterministe, alors une entrée dans la table de transition est un état unique.

2.2 Notre automate

Notre automate a pour but de transformer un texte en image, nous présenterons dans une première partie l'alphabet d'entrée que traite notre automate, puis les états que peut prendre l'automate, enfin nous verrons en détail les différentes transitions.

2.2.1 L'alphabet de l'automate

L'alphabet est l'ensemble des symboles que l'automate est susceptible de reconnaître et donc de traiter. Un texte peut être découpé en de nombreux symboles ou unités textuelles. Les unités textuelles les plus usuelles sont les mots et les n-grammes, mais bien d'autres unités textuelles existent et peuvent être utilisées. (Jalam et Chauchat 2002) apporte des arguments tendant à proposer les n-grammes plus que des mots. D'une manière générale, dans le domaine du texte, l'alphabet sera l'ensemble des unités textuelles possibles (tous les mots possibles, tous les n-grammes possibles, …). On notera UE cet ensemble. On définira les textes de la manière suivante : UE* ; c'est-à-dire l'ensemble de toutes les unités textuelles, y compris ε, la chaîne vide. Un texte est donc une succession finie d'unités textuelles.

De par sa composition, UE a une cardinalité assez forte. Du fait de cette cardinalité importante, il est difficile de trouver des transitions pour un tel nombre de symboles. Nous avons choisi de limiter notre alphabet à 25 symboles. Pour cela, on construit un alphabet relatif à chaque texte en entrée. Pour chaque texte, on crée donc 25 classes d'unités textuelles numérotées relativement à leur rang sur l'échelle des fréquences d'apparition dans le texte traité, les unités textuelles absentes n'étant, évidemment, pas prises en compte dans cette mise en classes. Ces 25 classes deviennent les 25 symboles de l'alphabet d'entrée de l'automate.

2.2.2 Etats de l'automate

Afin de présenter les états possibles que peut prendre l'automate, on définira dans un premier temps la notion d'image, puis on verra deux caractéristiques propres à la création d'images, enfin on verra la composition des états.

2.2.2.1 Définitions

Soit $I^{m,n}$ l'ensemble des images en niveaux de gris de taille m*n. $I^{m,n}$ a une cardinalité de 256^{m*n}. On peut donc noter :

$$I_{m,n}=\left\{I_{m,n}^{k}\right\}_{k\in\left[1,256^{m*n}\right]} \qquad (1)$$

où chaque image peut être considérée comme un ensemble connexe de points dans un espace à deux dimensions :

$$I_{m,n}^{k}=\left\{I_{m,n}^{k}(i,j)\right\} \quad \text{et} \quad i\in[1,m]\times j\in[1,n] \qquad (2)$$

et où chaque point peut prendre l'une des valeurs de niveaux de gris possibles :

$$I_{m,n}^{k}(i,j)\in[0,255] \qquad (3)$$

2.2.2.2 Les caractéristiques

Nous allons nous référer à un mode de formation des sons, suite temporelle d'informations, un signal monodimensionnel. Pour créer un son, il suffit de partir d'une courbe de silence, en général de valeur 0 sur toute la durée, et de modifier cette courbe d'une certaine intensité. La courbe, ainsi transformée crée un son.

Une image est un ensemble de points affectés d'un niveau de gris. Sur le même principe que le son, créer une image pourrait revenir à partir d'une image dite « silence », à modifier la surface de cette image en plus noir ou plus blanc. L'image dite « silence » serait une image dont tous les points seraient initialisés à un niveau de gris moyen

Alors que dans l'exemple du son, le temps fournit l'ordre de modification de la courbe, aucun chemin naturel ne permet de parcourir une image. Il existe certes des chemins de type fractal, par exemple la courbe de Peano (Nikolaou 2002), ils ne correspondent à aucun parcours naturel.

Dans notre application où l'objectif est l'étude d'un texte, il est nécessaire d'opter pour un parcours lié au contenu du texte. On peut définir un point dit de référence comme le point à partir duquel se fait la modification de l'image. De manière à conserver un maximum de degré de liberté aux déplacements, nous avons choisi le centre de l'image comme position « silence ».

2.2.2.3 Les états

L'image finale est construite par modifications successives. La modification se fait au point de référence, il est modifié à chaque opération.

Chaque état de l'automate est donc un 2-uplet contenant une image de dimension m*n en niveaux de gris et un vecteur de position (i,j) avec $1 \leq i \leq m$ et $1 \leq j \leq n$ donnant la position du point de référence.

Comme on l'a vu dans les rappels sur les images, il existe 256^{m*n} images en niveaux de gris de taille m*n possibles. Le vecteur position peut prendre m*n valeurs possibles dans une même image. On dénombre donc $m*n*256^{m*n}$ états possibles, par exemple 256 états si m=n=1 ou $1,7*10^{10}$ états possibles si m=n=2.

Vu que cet ensemble est fini, l'automate que nous construisons est lui aussi fini. L'état initial est le seul état défini par (image « silence », position « silence »), c'est-à-dire le 2-uplet constitué d'une image dont tous les points ont un niveau de gris moyen et d'une position de référence au centre de l'image. L'ensemble des états finaux est l'ensemble des états.

2.2.3 Transitions de l'automate

On a vu dans la section précédente que le nombre d'états, même s'il est dénombrable, est très important, c'est-à-dire qu'aucun graphique ne pourrait représenter la totalité de l'automate et que l'on ne peut pas créer une table des transitions. il n'est pas possible d'expliciter les tables de transitions qui listent tous les états possibles. Une telle table serait composée de $25*m*n*256^{m*n}$ lignes. En prenant m=n=1, les images constituées d'un seul point, la table aurait 6400 lignes.

Nous choisissons une description en compréhension en représentant tous les états plutôt qu'en les différenciant. Notre table des transitions est donc adaptée à notre problème.

Les transformations font intervenir des valeurs relatives aussi bien au niveau des positions que des niveaux de gris. Notre table de transitions présente donc les changements effectués sur ces paramètres. La position de référence est représentée par ses coordonnées X et Y, où X représente la position horizontale, X=0 correspond à la gauche de l'image et X=m correspond à la droite de l'image ; et où Y représente la position verticale, Y=0 correspond au bas de l'image et Y=n correspond au haut. Le niveau de gris est, quant à lui, représenté par NdG, où NdG=0 correspond à la valeur de niveau de gris la plus basse, le noir, et NdG=255 correspond à la valeur de niveau de gris la plus haute, le blanc.

La table 1 liste, selon le symbole en entrée, l'action à réaliser sur les paramètres de position et de niveau de gris.

Symbole D'entrée	Action réalisée
1	Rien
2	NdG=NdG +1
3	NdG=NdG –1
4	Y=Y+1
5	Y=Y-1
6	X=X+1
7	X=X-1
8	Y=Y+1 X=X+1
9	Y=Y+1 X=X-1

Symbole d'entrée	Action réalisée
10	Y=Y-1 X=X+1
11	Y=Y-1 X=X-1
12	NdG=NdD+5
13	NdG=NdG-5
14	NdG=NdG+7
15	NdG=NdG-7
16	NdG=NdG+10
17	NdG=NdG-10
18	Y=Y+2

Symbole d'entrée	Action réalisée
19	Y=Y-2
20	X=X+2
21	X=X-2
22	NdG=255
23	NdG=0
24	Y=n/2 X=m/2
25	Y=n/2-Y X=m/2-X

TAB 1 – *Table des transitions réalisées selon le symbole d'entrée.*

Cette table de transitions est très liée à la classification qui a été réalisée en prétraitement. Nos choix reposent sur les analyses du langage qui ont pu être faites par Zipf (Zipf 1935) et qui sont confirmées par ses expérimentations.

En effet, cette table indique que les 4% des unités textuelles les plus fréquentes n'ont aucune action associée.

Et d'une manière générale, plus les unités textuelles ont une fréquence d'apparition forte plus leur action de modification est faible, et plus les unités textuelles ont une fréquence d'apparition faible plus les modifications qui leur sont associées seront radicales, comme la mise à blanc ou à noir d'un point, la remise au centre de la position ou un changement de la position pour sa symétrie centrale.

Nous avons fait dépendre les états de l'automate de l'image et de la position de référence, notre automate est fini et déterministe.

2.2.4 Exemple

Voici un exemple de résultat obtenu sur un texte artificiel où les mots ont été placés de manière à montrer des résultats visuels sur les premiers états de l'automate.

Nous montrons donc les 5 premiers états, ainsi que l'état final.

La première transition effectue un assombrissement, la deuxième et la troisième effectuent des déplacements, tout d'abord vers la gauche, puis vers le bas, la quatrième transition effectue un éclaircissement, la cinquième déplace la position de référence vers le haut. L'image finale est composée de multiples points, dont les niveaux de gris n'est parfois que très légèrement modifié, on peut observer, entre autres, que le point de référence initial a à nouveau changé de niveau de gris. La position de référence est marquée par un cadre rouge.

FIG. 1 – *Etat initial*

FIG. 2 – *Etat 1 suite à un assombrissement*

FIG. *3 – Etat 2 suite à un déplacement vers la gauche*

FIG. 4 – *Etat 3 suite à un déplacement vers le bas*

FIG. 5 – Etat 4 suite à un éclaircissement

FIG. *6 – Etat final*

3 Méthode d'évaluation de la représentation

L'automate crée une chaîne d'états à partir du texte. Chaque chaîne a en commun son premier maillon puisqu'il correspond à l'état initial de l'automate. L'état final représente quant à lui la vue définitive globale du texte, le texte constitue une accumulation d'éléments. Il est normal de s'intéresser à un tel état, puisqu'il est la somme de toutes les informations, puisqu'il contient toutes les caractéristiques du texte. Mais on peut dire, en terme de classification textuelle ou de classification d'auteurs, chaque texte est unique parce que c'est le résultat final mais aussi grandement l'ensemble du processus de création.

En mathématiques, une propriété n'a de sens que si sa preuve en a aussi. Or qu'est-ce qu'une preuve, si ce n'est une suite logique de définitions agencées de manière logique de façon à rendre la propriété irréfutable.

Un texte est aussi une suite logique d'informations. Si deux textes restent neutres face à une prise de décision quelconque, l'un peut être neutre après avoir apporté autant d'arguments positifs que d'arguments négatifs et l'autre peut être neutre en n'ayant aucun avis sur le sujet. Il devient donc plus intéressant de prendre en compte l'évolution du texte que la conclusion finale.

L'automate crée une chaîne d'états à partir du texte. Cette chaîne peut être plus ou moins longue selon le texte. Comment évaluer d'une manière identique un ensemble aussi hétérogène de suites d'images. Nous avons fixé un nombre de mesures. Ce nombre N de mesures fixé, la chaîne d'état est divisée en N+1 sous chaînes équitables. On considère que la division d'une chaîne de longueur L est faite de manière équitable lorsque la longueur des sous chaînes est comprise dans l'intervalle ouvert :

$$\left]\frac{L}{N+1}-1;\frac{L}{N+1}+1\right[\quad (4)$$

Une fois ce découpage établi, on ressort les N+1 états obtenus à la fin des N+1 sous chaînes dans le séquençage de l'automate. Les N caractéristiques extraites sont les N distances entre les N+1 états successifs.

Un état représente essentiellement une image finie correspondant à une étape de réalisation du texte. Il est inutile de prendre en compte la position de référence qui n'a de sens qu'au niveau de la poursuite du texte.

Comme nous souhaitons évaluer l'évolution entre deux étapes de construction du texte, une distance entre deux états amène à une distance entre deux images. Plus précisément, nous considérons la différence de distorsions que l'image initiale a subies pour atteindre cet état, avec une importance plus grande à mesure que l'on se rapproche de la position initiale. Dans un système ayant pour centre la position de l'état initial , l'importance est gérée par le regroupement des caractéristiques des points ayant la même valeur pour valeur maximum de leur vecteur position.

Soient deux images $I_{m,n}^{k1}$ et $I_{m,n}^{k2}$ et soit la position de point de référence à l'état initial : $x_0=\frac{m}{2}; y_0=\frac{n}{2}; NdG_0=127$

la distance entre les deux images est donc :

$$D(I_{m,n}^{k1}, I_{m,n}^{k2}) = \| V_{k1}; V_{k\acute{e}} \| \text{ pour } d \in \left[0; \max(\frac{m}{2}, \frac{n}{2})\right] \quad (5)$$

$$V_{k1}(d) = \sum_{\max(x_0-d,m)}^{\min(x_0+d,m)} \left| NdG_0 - NdG_{k1}(x,d) \right| + \sum_{\max(y_0-d,n)}^{\min(y_0+d,n)} \left| NdG_0 - NdG_{k1}(d,y) \right| \quad (6)$$

et $NdG_{kl}(x,y)$ la valeur de niveau de gris du point de coordonnées (x,y) de l'image $I_{m,n}^{k1}$.

Il ressort ainsi de l'application de l'automate N valeurs caractérisant l'évolution de la création de l'image mais aussi de l'idée qui est construite du texte au cours de la lecture.

L'analyse finale des caractéristiques de l'évolution des textes est faite par une Analyse par Composantes Multiples. En fait, toute distance dans l'espace R^N pourrait être utilisée.

4 Présentations de divers résultats

Dans cette partie nous présentons deux analyses obtenues par notre méthode. Elles ont toutes les deux été paramétrées de la même manière, c'est-à-dire que l'alphabet d'entrée est construit sur des 4-grammes de caractères, l'image créée a pour taille m=n=501 afin de ne pas avoir d'effet de bords (paramètre fixé de manière expérimentale), et 100 distances sont calculées lors de la création de l'image. Pour chacune la première analyse, une classification hiérarchique présentera la formation des groupes afin de mieux illustrer l'analyse. Pour la seconde analyse, les explications trouveront plus de sens par une représentation graphique sur les deux premiers axes factoriels.

4.1 Analyse intra corpus

Cette première analyse a pour but de prouver que notre méthode trouve un sens lorsqu'on l'applique dans un contexte textuel en particulier. Le corpus étudié est l'ensemble des œuvres de Maupassant (Marteau et al. 2003), il est constitué de 8 textes de Maupassant : Une Vie (1882), Bel-Ami (1885), Mont-Oriol (1887), Pierre et Jean (1888), Fort Comme la Mort (1889), Notre Cœur (1890), L'Ame Etrangère (1890), L'Angélus (1891).

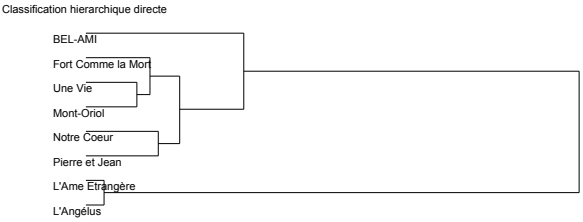

FIG. 7 – *Classification Hiérarchique des oeuvres de Maupassant*

N'oublions pas qu'à partir de 1884, Maupassant souffre de névralgies, est victime d'hallucinations et de crises d'angoisse, son mal se répand. Une syphilis, évoluant en paralysie générale, assombrit ses dernières années, endeuillées par la folie d'Hervé, son frère cadet qui meurt en novembre 1889. Les mois suivants, les phénomènes morbides s'aggravent et le 6 janvier 1892, après une tentative de suicide, il est interné à Passy, dans la clinique du docteur Blanche (où son frère avait été interné). Suivent dix-huit mois de souffrance, entrecoupés de brusques exaltations pendant lesquelles il affirmait communiquer avec l'au-delà. Cette fin d'existence végétative s'achève le 6 juillet 1893, laissant deux romans inachevés « l'Angélus » et « L'Ame étrangère », Maupassant n'aura jamais retrouvé la lucidité.

La figure 7 montre de manière évidente que les deux romans « l'Angélus » et « L'Ame étrangère » se distinguent parfaitement des autres romans.

4.2 Application à l'étude d'entretiens sociologiques

La méthode a été a priori développée pour traiter de manière automatique des entretiens sociologiques sur un thème donné, retranscrits sous forme de texte. Le but précis était de fournir un outil d'aide à la création de groupes sociologiques sur une thématique précise. La partie des entretiens étudiée ici est une partie traitant de la vie familiale, de l'amélioration ou de la dégradation de celle-ci par des éléments extérieurs, des enfants, etc. Ce corpus est constitué de 29 corpus dont 27 entretiens individuels et de 2 entretiens au cours desquels deux personnes interviennent. L'utilisation du langage oral rend difficile l'utilisation des méthodes classiques d'analyse de textes reposant sur une analyse linguistique du texte.

La figure 8 présente le résultat de l'analyse en composantes principales de la représentation des textes par le vecteur à 100 composantes dont nous avons décrit l'obtention dans la section 3. Les conclusions quant à la position de chacun des textes, en fonction d'un sens sociologique des deux axes factoriels apparaissent rapidement quand on s'intéresse aux personnes situées sur des positions extrêmes.

Dans un premier temps, on peut observer les individus 523 et 632 qui se trouvent à des positions opposées par rapport à l'axe factoriel 2. Ces deux individus ont un discours qui est complètement opposé. L'un envisage les choses de manière positive et l'autre a un sentiment de crainte et de méfiance vis-à-vis des effets sur la famille. L'individu 605 apporte autant d'avis positif que d'avis négatif, il trouve donc parfaitement sa place entre les deux individus précédents. Les individus 115 et 634, n'ont eu aucun avis en particulier puisqu'ils sont célibataires et n'ont pas de réel avis en ce qui concerne la vie de famille.

Le graphique permet donc de voir qu'un axe factoriel fait ressortir un point de vue global sur l'entretien et l'autre axe fait ressortir la construction, le raisonnement. Ainsi les individus 115, 634 et 605 qui, d'un point de vue global, peuvent paraître neutres puisque la somme des informations positives et négatives est neutre, se distinguent par leur raisonnement.

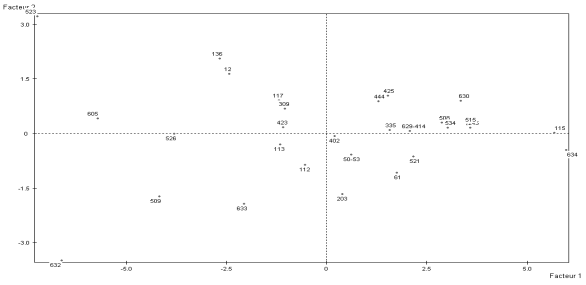

FIG. 8 – *Placement des individus par rapport aux deux premiers axes factoriels*

5 Conclusion

Nous proposons ici une méthode basée sur la construction d'un automate qui crée une image représentative du texte traité. Cette méthode laisse beaucoup de liberté dans l'ensemble des paramètres, que ce soient les paramètres de prétraitement ou les règles de transitions, la méthode d'évaluation ou la méthode de classification. Dans de prochaines études nous aborderons plus précisément l'influence des choix qui ont été faits, en particulier dans la table de transition.

Cet article présente des résultats montrant qu'utiliser une telle méthode a un véritable intérêt et un sens au point de vue de la classification sociologique. Nous avons pu la valider auprès des experts. Une telle représentation et l'étude approfondie de certains individus constitueraient une aide certaine lors de l'interprétation des enquêtes.

Références

Aho A., Sethi R. et Ullman J. (1986), Compilers, Addison Wesley, 1986.

Bavaud F. et Xanthos A. (2002), Thermodynamique et Statistique Textuelle, JADT 2002.

Benhadid I., Meunier J.-G., Hamidi S., Remaki Z. et Nyongua M. (1998), Etude expérimentale comparative des méthodes statistiques pour la classification des données textuelles, JADT 1998.

Benzécri J.P. (1973), L'analyse de données, Tome 1 et 2, Dunod Edition, 1973.

Biskri I. et Meunier J.-G., L'analyse de l'information multidimensionnelle au moyen des n-grams de caractères, Les Actes des 9es Journées Francophones d'Informatique Médicale, 2002.

Boufaden N., Lapalme G. et Bengio Y. (2001), Topic Segmentation : A first Stage to Dialog-Based Information Extraction, Natural - Language Processing Rim Symposium, NLPRS'01, pages 273–280, 2001.

Cavnar W.B. et Trenkle J.M. (1994), N-Gram Based Text Categorization, Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval", April 1994, pp 161-169.

Crowder G. et Nicholas C. (1997), Using Statistical Properties of Text to Create Metadata, SIGIR'97 Workshop on Network Information Retrieval, Philadelphia, August 1997.

Deschavanne P., Giron A., Vilain J., Fagot G. et Fertil B. (1999), Genomic signature: characterization and classification of species assessed by Chaos Game Representation of sequences, Mol. Biol. Evol. 16 (10): 1391-9, 1999.

Dufraigne C., Fertil B, Giron A et Deschavanne P (2001), Utilisation de la signature génomique pour la recherche de transferts horizontaux, JOBIM 2001, p 161-167, Toulouse , 2001.

Ferret O., Grau B., Minel J.-L. et Porhiel S. (2001), Repérage de structures thématiques dans des textes, TALN 2001.

Ferret O., Grau B. et Masson N. (1997), Utilisation d'un réseau de co-occurrences lexicales pour améliorer leur analyse thématique fondée sur la distribution des mots, 1ères Journées du Chapitre Français de l'ISKO, Lille, France, 1997.

Forest D. et Meunier J.-G. (2002), La Classification mathématique des textes, JADT 2002.

Garnier B. et Guérin-Pace F. (1998), La Statistique textuelle pour traiter les questions ouvertes, JADT 1998.

Gu Z. et Berleant D. (2000), Hash Table Size for Storing N-Grams for Text Processing, Technical Report 10-00a, Electrical and Computer Engineering, Ames, Iowa, 2000.

Hubert P., Labbé C., Labbé D. (2002), Segmentation automatique des corpus : Voyages de l'autre côté de J.M. Le Clézio, JADT 2002.

Jalam R. et Chauchat J.-H. (2002), Pourquoi les n-grammes permettent de classer des textes ? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques, JADT 2002.

Karthik M.N. et Davis M.(2002), Search Using N-gram Technique based statistical analysis for knowledge extraction in case reasoning systems, National Level Mathematics & CS Symposium, Chennai Mathematical Institute ICS '02, 2002.

Labbé D. (2003), Corneille dans l'ombre de Molière, Les Impressions Nouvelles, 2003.

Lelu A.(2002), Evaluation de trois mesures de similarité utilisées en science de l'information, Les journées d'étude des Systèmes d'Information Elaborée, 2002.

Lespinats S., Deschavanne P., Giron A. et Fertil B. (2004), Pertinence des métriques fractionnaires pour l'analyse des données de grande dimension (signatures génomiques), EGC 04, 2004.

Lespinats S., Deschavanne P., Giron A. et Fertil B. (2003), L'ADN en tant que texte : Style et Syntaxe. Une Syntaxe commune aux espèces, JDS 03, 2003.

Marteau H., Lefèvre A., Vincent N. (2003), Comparaison de textes par mesure fractale, Majestic 03, 2003.

Marteau H., Lefèvre A., Vincent N. (2004), Du texte à l'image, RFIA 04, 2004.

Morin A. (2002), Deux exemples d'analyse de données textuelles, Premier colloque sur la statistique et l'analyse des données dans les sciences appliquées et économiques, Beyrouth, Liban, 2002.
Mothe J., Christment C., Dkaki D., Dousset B. et Egret D.(2001), Information mining : Use of the document dimensions to analyse interactively a document set, European Colloquium on Information Retrieval Research, 2001.
Nikolaou N. et Papamarkos N. (2002), Color image retrieval using a fractal signature extraction technique, Engineering Applications of Artificial Intelligence, Vol. 15, p. 81-96, 2002.
Noureddine M. (1992), Théorie des langages, Office des publications universitaires, 1992.
Ouazzani I. (1998), Analyse Statistique des textes littéraires : l'exemple de Driss Chraibi, JADT 1998, 1998.
Pouliquen B (2002), Indexation de textes médicaux par extraction de concepts, et ses utilisations, Thèse de doctorat à l'Université de Rennes 1, 2002.
Reinheirt M. (1995), Quelques aspects du choix des unités d'analyse et de leur contrôle dans la méthode "Alceste", JADT 1995, 1995.
Reynar J.C. (1994), An Automatic Method of finding Topic Boundaries, Meeting of the Association for Computational Linguistics, 1994.
Salton G. (1971), The SMART Retrieval System - Experiments in automatic document processing, Prentice Hall Inc., Englewood Cliffs, NL, 1971.
Salton G. (1989), Automatic Text Processing, Addison Wesley, 1989.
SanJuan E. et Ibekwe-SanJuan F. (2002), Terminologie et classification automatique de textes, JADT 2002, St Malo, 2002.
Schbath S. (1997), An efficient statistic to detect over- and uder-represented words in DNA sequences, Journal of Computitional Biology, Volume 4, Number 2, 1997.
Schbath S., Prum B. et De Turckheim E. (1995), Exceptionnal Motifs in different Markov Chain models for a statistical analysis of DNA Sequences Journal of Computitional Biology, Volume 2, Number 3, 1995.
Singhal A. et Salton G. (1995), Automatic Text Browsing Using Vector Space Model, Proceedings of the 5th Dual-UseTechnologies and Applications Conference, 1995.
Zipf G.K. (1935), The Psychology of Language, an Introduction to Dynamic Philology, M.I.T. Press, Cambridge, Massachusetts, 1935.

Summary

There is no need to say that the textual corpus is such as it is unrelenting by people and what its growth confirms us the obligation using automatic tools of treatment. This article is more particularly interested in the characterization of texts and there even authors. Actually all the methods which exist work on one in flat of the treated texts.

We mean that these methods work on the finished document, without admitting that a progress exists between the beginning of the document and the end. We propose a method which tries to bring this notion of textual evolution by treating the text by a machine. This method aims at bringing additional information as for the logical progress adopted during the creation of the text. We shall present the rules of the machine and the chosen evaluation. Then, we shall present, among others, results, validated by experts, obtained on a corpus of sociological conversations.