

# The use of text mining methods to value a biostatistic work with a medical study on menopause

Bertrand Malet<sup>\*</sup>, Pierre Pompidor<sup>\*</sup>, Michel Sala<sup>\* et \*\*</sup>,  
Sophie Bouvet<sup>\*\*\*</sup>, et Jean-Pierre Daurès<sup>\*\*\*</sup>

<sup>\*</sup> LIRMM- Laboratoire Informatique et Robotique de Montpellier, 161 rue Ada, 34392 Montpellier cedex 5 /  
CNRS / Université Montpellier II

<sup>\*\*</sup> Université Montpellier I

<sup>\*\*\*</sup> BESPIM, Biostatistiques, Epidémiologie clinique, Santé Publique et Information Médicale, du Centre  
Hospitalier Universitaire de Nîmes, Place du Pr R. Debré 30029 Nîmes cedex 9

**Abstract:** This paper is related to the text mining field, which can help classical biostatistics methods. Although a lot of text mining studies that have been made on the medical field were related to the molecular biology, less of them are based on clinical papers. The main objective here is to define an information correlation between free texts and classical statistical data. This project thus combines text mining techniques and the calculation and analysis methods from the biostatistics field.

**Keywords:** Text mining, biostatistic, menopause

**Résumé:** Cet article se situe dans le domaine de la fouille de textes (text mining) qui peut apporter une importante valeur ajoutée aux méthodes classiques de biostatistique. Si de nombreuses études de fouille de textes qui ont été effectuées dans le domaine médical entrent dans la spécialité de la biologie moléculaire, moins fréquentes sont celles qui portent sur des comptes rendus cliniques. L'objectif ici est de faire émerger la corrélation d'informations issues de textes libres, et des variables qui ont permis des résultats biostatistiques. Ce projet associe donc les techniques de fouille de textes avec celles de calcul et d'analyse du domaine de la biostatistique.

**Mots clés:** Fouille de textes, biostatistique, ménopause.

## INTRODUCTION

Entre 2001 et 2005 s'est déroulée l'enquête ANNA (« Age Nouveau - Nouvelles Attitudes »). Elle a été réalisée par l'institut Théramex en collaboration avec l'OMS (« Organisation Mondiale de la Santé ») auprès de 5137 patientes en période de ménopause. L'objectif était de rendre compte du ressenti des patientes au cours de cette période. Cette enquête a déjà fait l'objet de trois rapports d'analyses statistiques, respectivement réalisés en 2004, 2006 et 2007. Ces rapports ont permis de décrire l'effet des Traitements Hormonaux Substitutifs (THS) et du temps sur les troubles liés à la ménopause. Ce travail a été réalisé grâce à des analyses de l'évolution de la qualité de vie en fonction des comportements envers les THS. Ces rapports ont aussi mis en évidence les facteurs explicatifs de la prise (2006) et de l'arrêt (2007) des THS. Les données qui ont été utilisées dans ces rapports, et qui ont été rassemblées dans une matrice de données, sont extraites de cinq auto-questionnaires remplis par les patientes lors de cette enquête.

Parallèlement à ces questionnaires, il a été demandé aux patientes qui l'ont souhaité, de donner de façon libre leur propre définition de la ménopause. Ces définitions, jusque là non exploitées, constituent une source d'informations supplémentaires potentielles pour les statisticiens en charge de cette étude. L'objectif de ce projet est donc de rechercher, via une méthodologie de fouille de textes, de nouvelles informations pouvant être mises en corrélation avec les données déjà existantes dans le but de les compléter.

Cet article va décrire toutes les grandes étapes qui ont été réalisées dans cet objectif. Nous détaillerons d'abord les différentes phases de normalisation du corpus, comme les pré-traitements formels, l'utilisation de l'algorithme de Martin Porter pour normaliser les mots, ou encore l'utilisation d'un dictionnaire de synonymes créé par nos soins à partir du projet de thésaurus libre français WOLF. Cette étape de normalisation du corpus a plusieurs objectifs: faciliter le traitement ultérieur des données, et faire émerger les concepts utilisés dans le corpus tout en réduisant les dimensions de la matrice de données

sortante.

Une fois cette normalisation du corpus terminée, nous nous attarderons sur la pondération des différents concepts pour chaque texte du corpus afin d'en déterminer les plus pertinents. Pour ce faire, la méthode utilisée est la classique méthode du TF/IDF (« Term Frequency / Inverse Document Frequency ») que nous avons adaptée à notre problématique.

Enfin, nous évoquerons la façon dont les différentes données, nouvelles et anciennes ont été combinées, et les traitements dont font actuellement l'objet, ces données sous la direction d'une équipe de spécialistes. Nous discuterons également sur divers traitements pouvant être mis en place, soit pour affiner les résultats qui seront obtenus si besoin est, soit pour effectuer d'autres types de recherches.

## 1. L'étude d'origine

On ne peut aborder pleinement l'analyse et le traitement des textes rédigés par les patientes, si l'on ne s'intéresse pas en premier lieu à l'enquête ANNA (« Age Nouveau - Nouvelles Attitudes ») elle-même. Nous allons d'abord présenter l'enquête elle-même. Nous aborderons ensuite les questionnaires constituant ANNA, avant de résumer les études déjà effectuées en se basant sur ceux-ci. Nous terminerons par la présentation formelle des textes qui seront utilisés pour l'analyse.

Les informations contenues dans les paragraphes 1.2, 1.3, et 1.4 ci-dessous sont principalement tirées de [BOU 07].

### 1.1. L'enquête ANNA

Le terme ménopause désigne la période dans la vie d'une femme, généralement aux alentours de la cinquantaine, qui correspond à l'arrêt de l'activité ovarienne. Cet arrêt a des conséquences, plus ou moins marquées et durables selon les individus. En général, l'arrêt de la sécrétion hormonale ne se fait pas de façon brutale, mais plutôt progressive. Aussi, on considère une femme comme étant ménopausée seulement douze mois après l'arrêt définitif de ses règles.

On distingue plusieurs types de manifestations possibles des gênes engendrées par la ménopause, comme l'ostéoporose, des manifestations vasomotrices (bouffées de chaleur, sudations nocturnes), des manifestations génitales (sécheresse et atrophie génitale), ou encore des manifestations urinaires (dysurie, troubles des émissions d'urine).

Parallèlement à ces troubles, d'autres symptômes sont subis par 30% à 50% des femmes, même si leur lien direct avec la ménopause n'a pas été établi, mais restent des facteurs de trouble important de la qualité de vie des femmes pendant leur ménopause. On note parmi eux des troubles psychosociaux (irritabilité, tristesse, trous de mémoire), des prises de poids, ou encore la fatigue.

On constate chez un quart des femmes la présence

de ces facteurs encore dix ans après la ménopause, même si leur fréquence d'apparition et leur sévérité diminuent avec le temps. C'est suite à la présence de tous ces troubles qu'un traitement hormonal substitutif (THS) a été mis en place dans le but de compenser les manques en oestrogènes.

L'enquête ANNA est une étude de cohorte de 5137 femmes nées en 1951 qui se déroule sur 5 ans, de 2001 à 2005. Ce large échantillon de femmes a été constitué en 2001 à la suite d'une campagne d'affichage dans les cabinets médicaux de médecins généralistes et de gynécologues, ainsi que par l'intermédiaire de 200 pharmacies de France, par diffusion dans les revues grand public, et également par le biais de média audio-visuel telles que des émissions de télévision et de radio. La mise en place d'un numéro vert a permis de recruter ces 5137 femmes volontaires pour être suivies pendant 5 ans. Parmi ces femmes, 3517 ont répondu aux cinq auto-questionnaires de l'enquête ANNA ce qui représente un taux de participation aux cinq enquêtes de 68.46%.

Le but de cette enquête est d'évaluer la qualité de vie des femmes en fonction de leur statut ménopausique, de leur comportement vis-à-vis des THS (« Traitement Hormonal de Substitution ») et du temps.

### 1.2. Les données collectées

L'enquête ANNA a bien entendu consisté en un rassemblement de connaissances et d'informations au sujet des différentes patientes. Ces dernières ont dû remplir un questionnaire annuel, tout au long de leur participation à l'enquête. Ces questionnaires sont au nombre de cinq :

- Le premier questionnaire, « ANNA0 » était constitué de 31 questions, et a été proposé aux femmes en 2001, soit au début de l'enquête.
- Le second questionnaire, « ANNA2 » comptait 48 questions et a été rempli en 2002.
- Le troisième questionnaire, « ANNA3 », rempli en 2003, contenait 51 questions.
- Les quatrième et cinquième questionnaires, « ANNA4 » et « ANNA5 », respectivement proposés en 2004 et 2005, comptaient 48 questions.

Ces questionnaires sont composés de plusieurs types de questions :

- Questions qualitatives ordinales (un peu, beaucoup...)
- Questions dichotomiques (oui, non)
- Questions quantitatives simples
- Questions de type d'échelles visuelles analogiques (EVA, échelle visuelle entre deux valeurs sur laquelle la personne remplissant le questionnaire barre l'endroit qu'elle considère comme répondant correctement à la question)

Les contenus de ces questionnaires ont été

regroupés en une matrice de données composée de 5137 lignes (une ligne par patiente) sur 1773 colonnes, correspondant aux variables les décrivant. On notera que parmi ces 1773 variables, toutes ne sont pas issues directement des questionnaires, mais ont été créées pour les besoins des analyses statistiques.

### 1.3. Les travaux réalisés

ANNA a déjà fait l'objet de trois rapports d'analyses statistiques, respectivement réalisés en 2004, 2006 et 2007. Ces rapports ont mis en évidence certains profils de patientes en fonction des comportements envers les THS et de données cliniques et administratives.

Les deux premiers rapports ont dû, en premier lieu, stabiliser la base de données en la rendant cohérente par des corrections. Ensuite, chacune des variables a été décrite et interprétée en fonction du statut ménopausique des femmes concernées. Les données ainsi rassemblées et normalisées entre elles ont alors été stabilisées afin de constituer une base exploitable par SAS, puissant logiciel de statistiques.

Le premier rapport, portant sur les résultats de suivi des trois premières enquêtes a mis en évidence le lien entre la ménopause et une période d'instabilité psychique et physique, mais a aussi montré que c'est un facteur nuisible significatif de la qualité de vie. Par ailleurs, ce rapport confirme l'action bénéfique du THS (« Traitement Hormonal de Substitution ») sur de nombreux troubles de la ménopause.

Le second, analysant les résultats des deux dernières enquêtes a confirmé les résultats des précédentes enquêtes en montrant encore l'action bénéfique du THS sur plusieurs troubles de la ménopause.

Le troisième rapport, reprend dans son ensemble l'étude des cinq questionnaires, et a pour objectifs: de mettre en évidence des liens entre la qualité de vie et le comportement thérapeutique envers les THS, puis de mettre en évidence des facteurs explicatifs de l'arrêt des THS. On distingue trois types de comportements thérapeutiques définissant trois groupes distincts de femmes:

- Les femmes qui prennent leur THS sans discontinuer.
- Les femmes ne prenant jamais de THS.
- Les femmes prenant un THS et qui l'arrêtent.

Ce rapport étudie l'évolution de la qualité de vie de chaque groupe, puis compare ces évolutions afin de définir l'effet du THS et celui du temps sur les troubles de la qualité de vie. Il définit également les facteurs pouvant expliquer les arrêts de traitement.

Il est important de comprendre ces rapports, et leurs relations, car les données qui seront obtenues à partir des textes qui nous intéressent, seront ajoutées à celles utilisées précédemment, afin d'affirmer ou d'infirmier ces thèses, voire en découvrir de nouvelles.

### 1.4. Les interviews textuelles

En parallèle aux questionnaires initiaux remplis par les patientes, on a demandé à ces dernières de donner leur propre définition de la ménopause. Ces définitions constituent une source potentielle d'informations supplémentaires encore inexploitées jusque-là.

Considérons maintenant les textes issus des interviews des patientes. Ils sont initialement au nombre de 455, mais après examens approfondis, seuls 430 de ces textes sont exploitables. Ensuite, au niveau de la forme, on peut constater pour chacun des textes deux parties distinctes:

- Tout d'abord un identifiant, composé d'une lettre, puis d'un nombre entier à trois chiffres. Cet identifiant de la personne ayant écrit le texte est le même que celui stocké dans la base de données relative aux questionnaires. Celui-ci permet le recoupement des données textuelles et des données cliniques de base ANNA.

- Le texte en lui-même, est de forme tout à fait quelconque. En effet, il a été demandé aux femmes de donner leur propre définition de la ménopause. En conséquence, on ne retrouve aucune ressemblance ni dans la forme ni dans le contenu. Certains textes sont simplement constitués d'un mot, d'autres sont écrits en vers, certains comportent des paragraphes entiers... On note que le plus grand texte du corpus ne dépasse pas la vingtaine de lignes.

## 2. La fouille de textes et ses premières applications en informatique médicale.

Avant de passer en revue les différentes étapes du projet, il convient de rappeler en quoi consiste la fouille de textes, et de citer quelques-unes de ses applications dans le domaine de l'informatique médicale.

A l'heure actuelle, on considère qu'environ 80% de l'information numérique est constituée d'informations textuelles. Le besoin d'être en mesure d'analyser le contenu de cette surcharge d'information s'est donc tout naturellement fait sentir.

La fouille de textes est une spécialisation de la fouille de données. Cette technique est souvent désignée sous l'anglicisme « Text Mining ». C'est un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de « nouveauté », ou de « similarité » dans des textes regroupés en corpus. Dans la pratique, cela revient à mettre en algorithmes un modèle simplifié des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques. Les disciplines impliquées sont donc principalement la linguistique calculatoire, l'ingénierie du langage, l'apprentissage artificiel, les statistiques et bien sûr l'informatique.

De plus, selon [IBE 07], l'utilisation de fouille de textes dans le domaine biomédical est très répandue de nos jours, principalement en raison de la richesse des

langages spécialisés du domaine, ainsi qu'en raison du grand nombre de publications dans ce domaine. L'auteur prend notamment l'exemple de la base bibliographique Medline, qui contient près de quinze millions de notices descriptives. D'après l'auteur, on trouve parmi les applications les plus fréquentes, l'identification de synonymes et des abréviations, la reconnaissance des entités nommées (ou REN), l'extraction des relations entre entités, ou encore la classification automatique de textes. Les travaux réalisés sont si nombreux qu'ils ont été listés dans [COH 05].

On peut également évoquer la conférence annuelle BioNLP qui se déroule à Columbus aux États-Unis et qui a essentiellement pour thème l'avancée de la recherche en fouille de textes dans le domaine de la biomédecine.

À titre d'exemples de la diversité et de la multitude des utilisations de techniques de fouilles de textes en biomédecine, nous pouvons citer ceux que donne [CHE 05]: l'auteur considère que la fouille de textes a été appliquée aux dossiers des patients et à d'autres documents cliniques pour faciliter la découverte d'informations. Le procédé est sensiblement le même que pour la fouille de textes en littérature. Par exemple, le système décrit dans [HAR 03] extrait des termes depuis des textes cliniques. Par ailleurs, le système MedLEE, développé par Friedman et Hripcsak en 1998, a été utilisé sur des textes de patients de forme libre. Il en a extrait des entités utiles pour identifier les patients ayant la tuberculose ou un cancer du sein. Ces conclusions sont basées sur l'admission des patients pour effectuer des radiologies de la poitrine, et leurs rapports de mammographies. Ces études ont été menées respectivement par Knirsch en 1999 et dans [JAI 97]. On retrouve une approche similaire dans [CHA 04], où la fouille de textes est utilisée pour détecter automatiquement la possibilité de maladies infectieuses chez les patients à partir des rapports cliniques les concernant.

Avec tous ces exemples à l'appui, l'utilité des techniques de fouille de textes dans le domaine médical n'est plus à démontrer, et on constate que les applications en sont toujours plus nombreuses et variées.

### 3. La normalisation du corpus

La première étape de notre projet concerne la normalisation des données textuelles. Ce processus de fouille de textes a deux objectifs: faciliter le traitement des données en les rendant de forme utilisable, et réduire les dimensions de la matrice de données sortante.

#### 3.1. Les Pré-traitements formels

Ces premiers traitements, même s'ils semblent triviaux, sont indispensables au bon déroulement des opérations futures. Ils concernent la mise en forme des données à proprement parler.

Dans un premier temps, il faut procéder au

nettoyage manuel du corpus. En effet, il est nécessaire de vérifier le corpus de textes afin d'en supprimer le plus possible les fautes d'orthographe, coquilles, et autres problèmes de mise en page. On évite également ainsi les problèmes d'identification des textes.

Une fois ces textes nettoyés, on passe à la seconde étape qui consiste à changer les lettres majuscules à l'intérieur des textes par leur équivalent en minuscules. On cherche ici à faciliter la reconnaissance des mots.

Enfin, on va supprimer purement et simplement tous les mots considérés comme vides de sens, également appelés « mots stop ». Les déterminants, superlatifs, verbes auxiliaires et autres expressions de la négation sont généralement considérés comme tels. Ils sont considérés comme non pertinents pour la simple et bonne raison qu'ils sont trop souvent employés et ne sont donc pas du tout représentatifs d'un texte.

La méthode de suppression est elle aussi simple. Une liste de ces mots est constituée. On recherche alors les itérations de ces mots à l'intérieur des textes.

#### 3.2. L'utilisation de l'algorithme de Porter pour normaliser les mots

Une fois les pré-traitements formels appliqués aux textes, le système va chercher à identifier les différentes formes des mots afin de les regrouper. Cette méthode s'appelle la racinisation (« stemmer » en anglais). Nous utilisons ici la fonction de racinisation du français implémentée dans l'API Lucene de la fondation Apache, distribuée sous licence Apache.

Cette fonction proposée par « Lucene » est basée sur l'algorithme de Martin Porter, originellement développé en 1979, dans le cadre d'un projet du Computer Laboratory de Cambridge. Cet algorithme avait pour fonction de supprimer les affixes d'un mot, de manière à obtenir la forme canonique du dit mot. Adapté à la langue française, cet algorithme est moins performant que l'original, à cause des plus grandes variations concernant les mots de la langue française, mais reste néanmoins une référence.

Voici en détails l'algorithme de Porter adapté à la langue française tel qu'il est employé dans « Lucene ». Avant toute chose, il est bon de rappeler que la langue française inclut les formes accentuées suivantes: « â », « à », « ç », « ë », « é », « ê », « è », « ï », « î », « ô », « û », « ù ». Et les lettres suivantes sont considérées comme les voyelles: « a », « e », « i », « o », « u », « y », « â », « à », « è », « é », « ê », « è », « ï », « î », « ô », « û », « ù ». Par ailleurs, on considère que les mots soumis à cet algorithme sont entièrement en minuscules, ce qui justifie d'autant plus les traitements précédents.

Une fois ces précisions effectuées, l'algorithme va effectuer quelques analyses sur le mot en question. On commence par passer en « I », « U » ou « Y » majuscule si la lettre minuscule correspondante est à la fois précédée et suivie par une autre voyelle, ou,

dans le cas de « u », s'il est précédé de la lettre « q ». Le but de cette opération est de ne pas classer les lettres passées en majuscules en tant que voyelles. Ensuite, le système découpe le mot en trois parties se chevauchant :

- On appelle RV la chaîne de caractères située après la première voyelle n'étant pas au début du mot, sauf si le mot commence par deux voyelles, auquel cas, on désignera comme RV, la séquence débutant à la troisième lettre du mot. Si aucune de ces positions ne parvient à être déterminée, on utilisera la fin du mot. Exceptionnellement, les chaînes de caractères « col », « par » et « tap », placées au commencement d'un mot peuvent également être sélectionnées, ainsi que la région située à leur droite, en tant que RV.

- On appelle R1 la chaîne de caractères située après la première lettre non-voyelle suivant une voyelle, ou la fin du mot s'il n'existe aucune non-voyelle de cette forme.

- On note également R2 la suite de lettres située après la première non-voyelle suivant une voyelle de R1, ou la fin du mot si on ne parvient pas à déterminer une telle lettre.

Le mot ainsi découpé, on peut le soumettre à l'algorithme lui-même qui va se dérouler en sept grandes étapes :

- La suppression des suffixes standards.

On recherche parmi une liste de suffixes le plus long suffixe contenu à la fin de notre mot, et on effectue un traitement correspondant à ce suffixe. Par exemple, pour les suffixes « atrice », « ateur », « ation », « atrices », « ateurs » ou « ations », on supprime le suffixe s'il est dans R2. Si par contre le suffixe est précédé par « ic », deux possibilités sont envisagées : on supprime le suffixe s'il est dans R2, sinon, on le remplace par « iqU ».

- La suppression des suffixes verbaux.

Les tests de cette étape ne sont effectués que sur la partie RV du mot et si aucun traitement n'a été réalisé à l'étape précédente, ou si l'un des suffixes « amment », « emment », « ment », « ments » a été trouvé.

Cette étape se déroule en parties bien distinctes : tout d'abord les suffixes commençant par « i », les autres ensuite. On n'effectue la seconde que si aucune modification n'a été apportée lors de la première. Comme lors de la première étape du traitement, on recherche le plus long des suffixes contenus dans notre mot parmi ceux d'une liste, et on effectue l'action correspondante.

- Si le mot a été modifié depuis le début de l'algorithme, on remplace un « Y » final par « i », et un « ç » final par un « c ».

- La suppression de suffixes résiduels.

Comme pour la suppression des suffixes verbaux, on recherche dans la partie RV du mot le plus long suffixe restant parmi une nouvelle liste. Par exemple,

pour les suffixes : « ier », « ière », « Ier », « Ière », on remplace le suffixe par « i ».

- Supprimer les lettres doubles.

Si le mot se termine par « enn », « onn », « ett », « ell » ou « eill », on supprime la dernière lettre.

- Suppression d'accents.

Si le mot se termine par « é » ou « è », suivi de le dernier caractère non-voyelle, on supprime l'accent du « e ».

- Enfin, on remplace les « I », « U » et « Y » restants par leur équivalent en minuscules.

Après avoir été soumis à cet algorithme, il ne reste d'un mot que la partie que l'on appelle sa racine. Par exemple, les formes verbales « (je) découvre » ou encore « (nous) découvrons » deviennent « découv ». On a donc ainsi un regroupement de toutes les déclinaisons d'un mot en une seule forme, limitant ainsi la multiplication des vecteurs (mots) de même sens.

### 3.3. L'utilisation d'un dictionnaire de synonymes

Une fois toutes les déclinaisons d'un même mot regroupées en une seule, il nous reste la possibilité de poursuivre cette action de regroupement en l'appliquant sur des mots différents, mais de sens voisin.

Nous avons conçu, à partir du projet thésaurus français libre WOLF, un dictionnaire de synonymes. Un thésaurus est une sorte de dictionnaire hiérarchisé. Il comporte non seulement des mots et leur définition, mais principalement d'autres mots, en relation avec les premiers, les contextes d'utilisations...

Le WOLF (« Wordnet Libre du Français ») est une ressource lexicale sémantique libre pour le français. C'est un projet développé par l'équipe Alpage de l'INRIA (Institut National de Recherche en Informatique et en Automatique). L'objectif de ce projet est de franciser, et avec un outil libre, le thésaurus de la langue anglaise WordNet de Princeton.

Le thésaurus WordNet a pour objectif de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Il regroupe des concepts en synsets (pour « synonym set »). Chaque synset contient les différents mots synonymes pouvant prendre ce sens, une définition du sens, les usages... Ce sont au total plus de 200 000 sens qui y sont recensés dans sa dernière version.

Les développeurs de WOLF ont choisi, plutôt que de recréer un maillage de sens à partir de langue française, de reprendre le maillage des synsets du WordNet anglais, et ce, dans le but de faciliter les conversions futures. On retrouve donc dans WOLF des synsets correspondants à des termes ou expressions anglaises n'ayant pas de traduction francophone, et ne contenant donc aucun mot. Les définitions et autres ressources ont par contre été

conservées. La version de WOLF utilisée ici (la dernière disponible) est la 0.1.4. Elle contient 115424 synsets différents.

Au niveau de la forme, WOLF se présente comme un fichier XML (« Extensible Markup Language ») qui est donc utilisable facilement par n'importe quel langage informatique. Chaque synset est donc logiquement intégré entre deux balises du même nom, ainsi que chacune de ses composantes. Les balises qui nous intéressent pour ce projet sont les balises « <LITTERAL> » au sein desquelles sont contenus chaque synonyme et son sens.

La constitution d'un dictionnaire fiable est une opération complexe et délicate. Elle s'est réalisée dans notre cas en plusieurs étapes bien distinctes:

- Tout d'abord, la solution logicielle parcourt le fichier XML de WOLF, afin de sélectionner les synsets jugés pertinents. On considère comme pertinents les synsets contenant au minimum deux termes synonymes. En effet, la plupart des synsets ne contiennent qu'un seul terme, voire aucun pour ceux n'ayant pas d'équivalence en français. Cette première version épurée de Wolf est alors sauvegardée par notre programme, et c'est maintenant à partir d'elle que l'on va travailler.

- Ensuite, on parcourt manuellement le fichier épuré afin de supprimer les termes dont l'utilisation dans Wolf ne correspond pas au sens principal, ainsi qu'un regroupement des mots issus de synsets de sens voisins, afin d'éviter tout doublon. On en profite également pour supprimer les termes latins et les noms propres.

- L'algorithme de lemmatisation de Lucene ayant ses limites pour la langue française, surtout au niveau de la lemmatisation des formes verbales, on ajoute de nouveaux synsets au thésaurus. Ces synsets remplacent les formes verbales non traitées par le même verbe à l'infinitif.

- Dans le même principe, on va ajouter des synsets contenant non pas des mots, mais des expressions usuelles rencontrées dans les fiches, visant à les remplacer soit par un mot ayant plus de sens, soit simplement de rassembler des expressions de sens voisin en une seule.

La plupart de ces modifications ont été réalisées manuellement, en modifiant directement le contenu du fichier XML correspondant à la version épurée de Wolf. Suite à l'ensemble de ces modifications, le nombre de synsets passe de 115424 pour le Wolf d'origine, à 4868 synsets dans sa version définitive qui sera utilisée par notre solution logicielle.

Une fois ce thésaurus constitué, on l'utilise de la manière suivante:

- On parcourt pour chaque interview, chacune des lignes de texte.

- Pour chaque ligne, on ajoute un espace en début et en fin de ligne, et on parcourt chaque synset du thésaurus.

- Pour chaque synset, si on trouve dans la ligne de texte l'un des synonymes (excepté le premier du synset) encadré par deux espaces, on remplace ce mot dans la ligne de texte par le premier synonyme du synset.

Cette étape dans notre processus a pour fonction de réduire le nombre de mots employés, en fusionnant les mots ayant un sens commun. On cherche ainsi à obtenir une matrice de concepts plutôt qu'une matrice de mots. En plus de réduire la dimension des données, le fait que le même intitulé désigne tous les mots de même sens améliore la cohérence des données extraites.

#### 4. L'utilisation d'un seuil de pondération de mots par un filtre basé sur la méthode TF/IDF.

L'ensemble des traitements appliqués lors de la normalisation du corpus ont permis de faire émerger un ensemble de mots plutôt que des phrases. Ces mots représentent les concepts présents dans le texte initial. On peut donc définir chaque texte de notre corpus comme un vecteur dont les composantes sont les concepts présents dans ce texte. Cependant, le fait qu'un concept soit ou non présent dans un texte est insuffisant pour définir pleinement le vecteur de ce texte. Il faut également attribuer une valeur, ou poids, à chacune des composantes de ce vecteur.

Pour se faire, nous avons mis en place un calcul de pondération de chaque mot pour chaque terme du corpus, afin de déterminer l'importance relative de l'utilisation de ce concept dans le texte en question. La formule que nous avons employée est la classique formule du TF/IDF (« Term Frequency / Inverse Document Frequency »). Cette fonction a été pour la première fois développée dans son ensemble dans Salton [SAL 73].

Cette mesure statistique permet d'évaluer l'importance d'un mot par rapport à un document extrait d'une collection ou d'un corpus. Le poids augmente proportionnellement en fonction du nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus. Des variantes de la formule originale sont souvent utilisées dans des moteurs de recherche pour apprécier la pertinence d'un document en fonction des critères de recherche de l'utilisateur. Voici néanmoins la plus fréquente et la plus répandue, décomposée:

- $tf_{i,j}$ : La fréquence du terme:

Il s'agit du nombre d'occurrences de ce terme dans le document considéré appartenant au corpus. Cette fréquence peut être éventuellement normalisée pour éviter les biais liés à la longueur du document (le nombre d'occurrences serait potentiellement plus élevé dans une page que dans un paragraphe). Cependant, dans notre cas, cette normalisation de la fréquence a été inutile du fait de la longueur relativement comparable des textes du corpus entre eux (de un mot à une vingtaine de lignes).

-  $idf_i$ : La fréquence inverse de documents telle que:

$$idf_i = \log \frac{N}{df_i} \quad (1)$$

Dans cette formule, on note:

-  $N$ : Le nombre de documents du corpus.

-  $df_i$ : Le nombre de documents contenant le terme  $i$ .

- La formule de base du TF/IDF est:

$$w_{i,j} = tf_{i,j} \cdot idf_i \quad (2)$$

On en déduit donc la formule intégrale:

$$w_{i,j} = tf_{i,j} \cdot \log \frac{N}{df_i} \quad (3)$$

Il existe de nombreuses variantes de cette formule utilisée dans beaucoup de projets basés sur des méthodes de fouille de textes. Selon les cas, certains auteurs des travaux en question ont normalisé la fréquence du terme par la taille du document, comme nous l'avons vu plus haut. Dans d'autres cas, les auteurs ont préféré normaliser la fréquence du mot par la fréquence maximale de tous les mots du document en question. C'est notamment la méthode préconisée dans [IBE 07].

Le fait d'appliquer cette formule du TF/IDF aux mots traités de notre corpus fournit ainsi la pertinence des termes pour chacun des documents dans lequel il apparaît. En effet, plus la valeur du poids affecté à un mot est importante par rapport à un texte donné, plus on peut considérer ce mot comme caractéristique du texte en question.

## 5. Mise en commun des données et traitements statistiques

### 5.1. La jonction des deux jeux de données

L'intérêt des données que nous avons extraites de ce corpus de textes réside dans le fait qu'elles vont être utilisées conjointement avec les informations issues des questionnaires de l'enquête ANNA. Il en résulte un croisement des deux matrices.

En effet, parmi les 5137 patientes ayant participé à l'enquête, on a extrait les informations relatives aux 430 dont nous avons pu exploiter les données textuelles. A ces données d'origine, on ajoute la matrice des données textuelles extraites par notre système. On obtient ainsi une nouvelle matrice de 430

patientes caractérisées par 3762 variables (1773 variables administratives et cliniques + 1989 concepts). On comprend ainsi l'intérêt de notre étude pour exploiter les textes qui ne l'étaient pas jusqu'à présent.

Cette matrice de données a été importée dans le logiciel SAS (« Statistical Analysis System »). C'est avec ce logiciel propriétaire que les différents traitements statistiques vont être effectués.

### 5.2. La représentativité des mots

A partir de la base de données de 1989 mots caractérisant les 430 interviews, une sélection des mots les plus représentatifs a été effectuée. Cette sélection s'est faite en deux temps :

Minimum	0.17
Médiane	1.56
Maximum	55.45
Moyenne	1.67
Écart-type	1.03

**Tableau 1 :** Quelques statistiques élémentaires sur les poids TF/IDF obtenus

- Tout d'abord par recodage de la base de données par rapport à la médiane du TFIDF de la matrice.

Si le TFIDF du mot  $i$  pour l'interview  $j$  est supérieur ou égal à la médiane ( $Q2 = 1.56$ ) alors la modalité vaut 1. Sinon elle vaut 0.

- Ensuite, les mots sont conservés si il y a au moins 2% de modalité 1 (c'est-à-dire  $TFIDF \geq 1.56$ ) pour le mot  $i$ .

On a ainsi obtenu une base de données à 430 interviews et 243 mots représentatifs de la question posée.

### 5.3. Formation de groupes de mots

#### 5.3.1. Analyse en correspondances multiples

Un individu est représenté par un point dans un espace à  $p$  dimensions ou  $p$  est le nombre de modalités. L'ensemble des individus constitue alors un nuage de points dans cet espace à  $p$  dimensions.

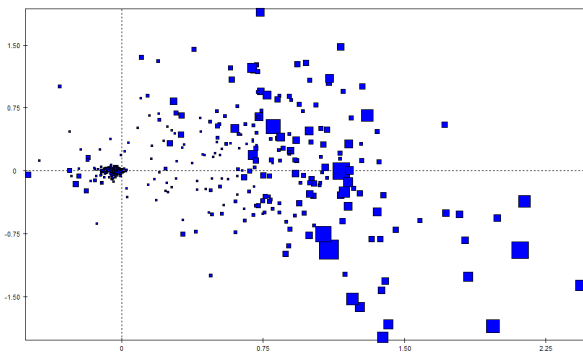
L'ACM permet de réduire la dimension de cet espace en construisant des axes qui sont des combinaisons linéaires des axes initiaux. Ces axes sont construits de telle manière que le premier axe explique le plus de variabilité des données, le second explique le plus de variabilité du résidu non expliqué par le premier, et ainsi de suite.

A chaque axe factoriel est associé un taux d'inertie qui permet d'exprimer la part de variance expliquée par cet axe.

Dans notre cas, les individus seraient complètement décrits par un espace à 486 dimensions où chaque axe expliquerait 0.2% de la variance totale du nuage de points. L'ACM réalisée permet la

réduction de l'espace à un plan, constitué par les deux premiers axes factoriels, qui explique 5.6% de l'inertie totale.

La représentation graphique ci-dessous explique donc 5.6% de la variance globale du nuage de points.



**Graphique 1:** Représentation graphique du plan 1-2 de l'ACM

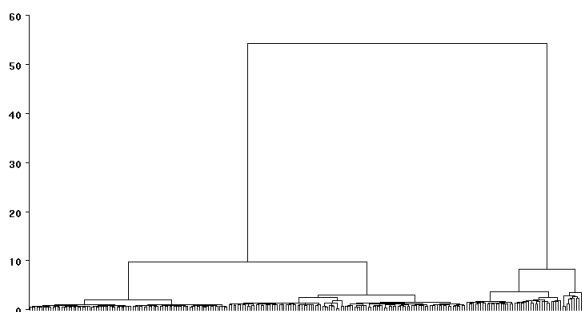
Il paraissait difficile de constituer des groupes de mots à partir de cette représentation graphique. D'autres méthodes ont donc été employées.

**5.3.2. Classification hiérarchique ascendante**

Cette méthode permet le regroupement successif de points par ordre de proximité décroissante. La distance entre les points peut être calculée de plusieurs façons. Ce sont neuf méthodes de calcul de distance différentes qui ont été utilisées : lien moyen, centroid, « complete linkage », « EML », bêta flexible, « McQuitty's Similarity Analysis », médiane, « single method » et ward.

La méthode donnant les meilleurs résultats est celle proposée par Lance, G. N. et Williams, W. T. (1967). « A General Theory of Classification Sorting Strategies. 1. Hierarchical Systems » *Computer Journal* n°9, pp. 373-380 et évaluée par Milligan, G. (1989). « A study of the beta-flexible clustering method », *Multivariate Behavioral Research*, n°24, pp. 163-176) qui offre l'avantage de ne pas favoriser l'émergence de groupes de même taille comme l'algorithme de Ward et qui permet de choisir un coefficient de lissage des données.

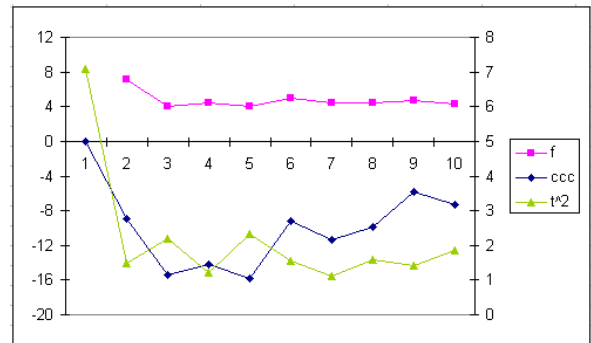
La valeur de bêta est fixée à -0.5 pour les meilleurs résultats. Le dendrogramme obtenu pour cette méthode et avec ce coefficient est tracé ci-dessous.



**Graphique 2:** Dendrogramme représentant la distance entre les clusters en fonctions des mots

Des critères objectifs sont calculés et conçus pour

présenter un maximum local au voisinage du nombre optimal de clusters. La représentation graphique ci-dessous permet de déterminer un nombre de cluster idéal.



**Graphique 3:** Indices permettant de déterminer un nombre optimum de clusters correspondant à un maximum local des courbes.

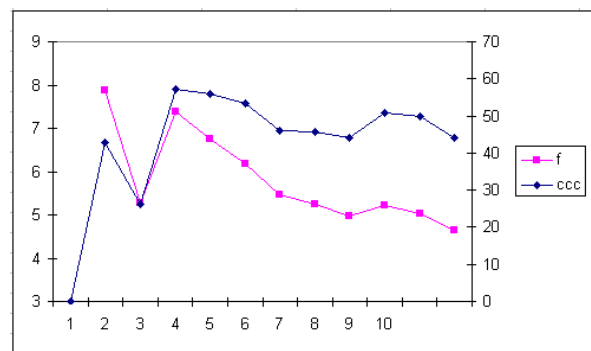
On peut remarquer que les paramètres du CCC et du pseudo F montrent des maximum locaux pour n = 4, n = 6 et n = 9 alors que le t<sup>2</sup> montrent des maximums locaux pour n = 3, n = 5 et n = 8.

La valeur n = 4 semble la plus proche d'un consensus et la plus compatible avec la structure du dendrogramme.

**5.3.3. Classification non hiérarchique**

C'est la méthode des nuées dynamiques. Son avantage est la simplicité et la rapidité des calculs à mettre en œuvre mais son inconvénient est le résultat du découpage souvent grossier et choix du nombre de clusters à priori.

Les critères du CCC et du pseudo F pour le choix du nombre de clusters ont été calculés (graphique 4).



**Graphique 4:** Indices permettant de déterminer un nombre optimum de clusters correspondant à un maximum local des courbes.

Cette méthode identifie comme optimal un nombre de 4 clusters comme pour la méthode précédente.

**5.4. Résultats des groupes de mots**

La classification hiérarchique permet la constitution d'un groupe de 87 mots, un de 42 mots, un de 9 mots et enfin un de 109 mots.

Avec la même base de données que pour la méthode hiérarchique, la méthode non hiérarchique

créé un groupe de 231 mots, un de 9 mots et de deux groupes de un mot chacun.

En enlevant successivement les mots isolés on pourrait obtenir deux groupes ; un de 8 mots et un de 231 mots.

Cependant les groupes obtenus ne sont pas identiques. A la lecture des groupes de mots et après réflexion, il est apparu que l'analyse devrait être reconduite sur les groupes nominaux. En effet, les mots seuls sont pour certains non informatifs (ex : « faire ») et l'absence des relevés de négations ne peut pas permettre d'interprétation claire des résultats.

## 6. Conclusion et ouverture

Dans un domaine où peu d'études avaient été faites, l'enquête ANNA a été accueillie très favorablement par l'ensemble des patientes qui y ont participé. Aujourd'hui, soit plus de trois ans après la fin de l'enquête, celle-ci n'a toujours pas livré toutes les informations cachées qu'elle contient. Il est cependant évident, que les données textuelles font partie de ces sources d'informations pas toujours exploitées.

Dans le cas présent, les techniques de fouille de textes ont permis de faire émerger de nombreuses données potentielles, qui sont, aujourd'hui encore, en cours d'évaluation et d'interprétation. Cependant, on peut d'ores et déjà imaginer quelques-unes des suites à apporter à ce projet.

Premièrement, dans le cas où l'analyse statistique des données extraites selon le procédé que nous venons de détailler, ne mettrait aucune corrélation, ou profil, en avant, il est envisagé de reprendre le travail en s'attardant cette fois non pas aux mots, mais aux entités nominales. Cette opération nécessiterait une analyse morpho-syntaxique du corpus, dans le but de reconnaître des groupements de termes. On ne travaillerait donc plus sur des mots mais sur des ensembles de mots groupés autour d'un nom.

Une seconde possibilité aujourd'hui envisagée, serait d'appliquer au corpus de textes, une ou plusieurs méthodes prédictives de catégorisation de textes. L'objectif serait de retrouver le classement de patientes réalisé a priori lors de l'exploitation des précédents travaux (c'est-à-dire les femmes prenant sans discontinuer un THS, celles n'en ayant jamais pris, et enfin celles ayant suivi un THS mais l'ayant arrêté) en utilisant uniquement les données contenues dans les définitions de la ménopause par ces même patientes.

Enfin, on peut imaginer prendre en compte, lors d'un futur développement, l'axe temporel des données. En effet, il pourrait être pertinent de comparer le contenu des données textuelles et des données cliniques au moment où l'on a demandé à la patiente de donner sa définition.

Ces diverses pistes sont encore à l'état d'étude, et nous évaluons encore laquelle nous allons privilégier

par la suite.

## BIBLIOGRAPHIE ET WEBOGRAPHIE

- [BOU 07] Bouvet S., 2007, *Analyse de la qualité de vie des femmes ménopausées en fonction de leur comportement thérapeutique vis-à-vis des Traitements Hormonaux Substitutifs*, Mémoire de stage de Master 2 en Biostatistique, Université de Montpellier II, 56p.
- [CHA 04] Chapman W. W., Dowling J. N., and Wagner M. M., 2004, *Fever Detection from Freetext Clinical Records for Biosurveillance*, Journal of Biomedical Informatics, 37, 120-127.
- [CHE 05] Chen H., Fuller S., Friedman C., and Hersh W., 2005, *Medical informatics, knowledge management and data mining in biomedicine*.
- [COH 05] Cohen A., Hersh W.R., 2005, *A survey of current work in biomedical text mining*.
- [FRI 01] Friedman C., Kra P., Yu H., 2001, Krauthammer M., and Rzhetsky A., *GENIES: A Natural-language Processing System for the Extraction of Molecular Pathways from Journal Articles*, Bioinformatics, 17(Supp. 1), S74-S82.
- [HAR 03] Harris M. R., Savova G. K., Johnson T. M., and Chute C. G., 2003, *A Term Extraction Tool for Expanding Content in the Domain of Functioning, Disability, and Health: Proof of Concept*, Journal of Biomedical Informatics, 36, 250-259.
- [KNI 96] Knirsch C.A., Jain N. L., Pablos-Mendez A., Friedman C., and Hripcsak G., 1996, *Respiratory Isolation of Tuberculosis Patients Using Clinical Guidelines and an Automated Clinical Decision Support System*, *Infection Control and Hospital Epidemiology*, 19(2), 94-100.
- [IBE 07] Ibekwe-SanJuan F., 2007, *Fouille de textes, méthodes, outils et applications*.
- [JAI 97] Jain N. L., and Friedman C., 1997, *Identification of Findings Suspicious for Breast Cancer Based on Natural Language Processing of Mammogram Reports*, in *Proceedings of the Fall 1997, AMIA Conference*, Philadelphia, USA, 829-833.
- [SALTON 73] Salton G., Yang C., 1973, *On the specification of term values in automatic indexing*.
- [http://lucene.apache.org/java/2\\_3\\_1/](http://lucene.apache.org/java/2_3_1/)
- <http://alpage.inria.fr/~sagot/wolf.html>