Espaces intrinsèques des relations entre mots : une exploration multi-échelle

Alain Lelu[1,2], Azim Roussanaly[1]

[1] LORIA, Nancy – azim.roussanaly@univ-lorraine.fr
[2] Université de Franche-Comté – alain.lelu@univ-fcomte.fr

Abstract

To determine the relationship of co-occurrence between words in a set of texts requires the selection of a span, i.e. a cutting into statistic entities of various size : from the plain N-gram (sliding span of N words) to full text through sub-sentence, sentence, paragraph, etc. These links can lead to various categorizations of words, depending on the "focus" used. Our study focuses on a corpus of newspaper articles (3 months of controversy about GMOs and endocrine disruptors) to which we apply 1) our Morph procedure for morpho-syntactic tagging, so as to disambiguate, tag and lemmatize to the best the sequence of forms, 2) our link validation test by multiple randomization of the presence matrix of tagged lemmas in text units of the chosen level, 3) our procedure for determining the intrinsic dimension of the matrix, which results in an estimate of the number of relevant clusters for each level of data granularity. Our results show that aggregated levels detect "stories" present in the corpus, that intermediate coarseness levels detect styles first, then collocations, with variable coagulation degree. This approach 1) generalizes the unsupervised labeling by (Schütze et al., 1995), based on N-grams of words, 2) determines the optimal space for representing words and chosen text units, *ie* that of the K* first non-trivial factors of correspondence analysis of the (binary, so far) matrix, where K* is determined by a randomization test, suitable for any distributions of rows and columns margins.

Résumé

Déterminer les liens de co-occurrence entre les mots d'un ensemble de textes nécessite le choix d'un empan, c'est à dire d'un découpage en individus statistiques de plus ou moins grande taille : depuis le simple N-gramme (empan glissant de N mots) jusqu'au texte complet, en passant par le virgulot, la phrase, le paragraphe, etc. Ces liens peuvent donner lieu à diverses catégorisations des mots, selon la "focale" utilisée. Notre étude porte sur un corpus d'articles de presse (3 mois de controverses sur les OGM et les perturbateurs endocriniens) auquel nous appliquons 1) notre procédure Morph d'étiquetage morpho-syntactique, de façon à désambiguïser, étiqueter et lemmatiser au mieux la séquence des formes présentes, 2) notre test de validation des liens, par randomisations multiples de la matrice de présence des lemmes étiquetés dans les unités textuelles du niveau choisi, 3) notre procédure de détermination de la dimension intrinsèque de cette matrice, dont découle une estimation du nombre de *clusters* pertinents pour chaque niveau de granularité de l'analyse. Nos résultats montrent que les niveaux les plus grands détectent les "histoires" dont il est question dans le corpus, ceux de grain intermédiaire détectent en premier lieu les styles, puis les collocations, de degré de figement plus ou moins important. Cette approche 1) généralise celle de l'étiquetage non-supervisé de (Schütze et al., 1995), basée sur les N-grammes de mots, 2) détermine l'espace de représentation optimal des mots et des unités de texte choisies, *i.e.* celui des K* premiers facteurs non-triviaux d'analyse factorielle des correspondances de la matrice (binaire, jusqu'ici), où K* est déterminé par un test de randomisation, adapté à n'importe quelle répartition des effectifs en lignes et en colonnes.

Mots-clés : Analyse des données textuelles, test de randomisation, validation de liens, validation de valeurs propres, Analyse Factorielle des Correspondances, laplacien de graphe, espace intrinsèque de matrice binaire.

1. Introduction : cadre d'analyse et objectifs

Déterminer les liens issus de la cooccurrence entre les mots d'un ensemble de textes nécessite le choix d'un empan, c'est à dire d'un découpage en individus statistiques de plus ou moins grande taille : depuis le simple N-gramme (empan glissant de N mots) jusqu'au texte complet,

en passant par le virgulot, la phrase, le paragraphe, etc. Ces liens peuvent donner lieu à diverses catégorisations des mots, selon la "focale" utilisée. A l'extrémité N-gramme cette approche est celle de l'étiquetage non-supervisé de (Schütze et al., 1995), basée sur les trigrammes de mots, qui aboutit à induire de façon non-supervisée les catégories syntaxiques présentes dans des textes de toute langue, connue ou inconnue. Nous tentons de généraliser cette approche en l'ancrant dans le cadre que nous avons désigné sous le terme d'*espace intrinsèque* d'une matrice (binaire, à ce jour) : on sait (Meila et Shi, 2000) que le partitionnement spectral de graphe consiste à extraire k *clusters* dans les k dimensions n°2 à n°k+1 de l'espace réduit défini par la décomposition aux valeurs singulières (SVD) de la « matrice laplacienne » de ce graphe, décomposition dont nous rappellerons les liens avec l'Analyse Factorielle des Correspondances (AFC) de sa matrice d'adjacence. Nous avons montré (Lelu et Cadot, 2011) qu'un test de randomisation permettait de déterminer le nombre K* de valeurs propres statistiquement significatives, que nous avons vérifié empiriquement être optimales pour des tâches d'apprentissage supervisé sur plusieurs jeux de données publics de partitionnement de graphes. Après avoir appliqué ce test à la détermination des liens (et anti-liens) significatifs dans le graphe des mots d'un corpus de dépêches Reuters (Lelu et Cadot, 2010), nous montrons ici 1) que cette approche peut être étendue à l'analyse de toute matrice binaire, 2) que dans le cas de matrices textes × mots elle peut donner lieu à une exploration multi-échelle des relations entre mots, en faisant varier le découpage des textes, le nombre de dimensions et de *clusters* extraits dans ces dimensions. Cette exploration, jamais conduite systématiquement à notre connaissance, et seulement effleurée dans le travail présenté ici, est susceptible de faire remonter à partir des seuls textes diverses dimensions du discours : citons, sans prétendre à l'exhaustivité, les rôles syntaxiques, les styles, les figements d'expressions, les relations sémantiques à plus ou moins longue portée… Beaucoup d'applications actuelles de l'analyse à grande échelle des textes sont en demande de procédures automatisées pour catégoriser certains de ces aspects, ou du moins assister un travail humain de catégorisation. Par exemple, dans une optique d'analyse à grande échelle d'articles scientifiques, il serait souhaitable de distinguer les mots descriptifs des matériels et méthodes utilisées, souvent transversaux à plusieurs domaines scientifiques, de ceux qui constituent le cœur de la problématique spécifique d'un front de recherche (« typer » les termes). Citons aussi l'assistance à la constitution d'ontologies, constitution pour laquelle le temps humain d'expertise est un facteur limitant.

2. Travaux proches

Nous avons listé dans la section I des approches heuristiques pour déterminer la dimensionnalité pertinente d'une matrice de données ; dans (Lelu, 2010) nous avons présenté un test dans le même esprit que celui développé ici : nous avons comparé les valeurs singulières d'une matrice binaire brute à ses contreparties issues de versions randomisées de cette matrice. Cependant cette approche est sujette à un souci statistique majeur : alors que l'éboulis des valeurs singulières traverse bien la borne haute de l'intervalle de confiance des valeurs singulières des matrices randomisées, définissant ainsi l'espace propre pertinent désiré, il traverse également la borne inférieure, créant un difficile problème d'interprétation pour les valeurs singulières "significativement petites". De plus cette approche n'a pas de lien avec les espaces propres Laplaciens, ni l'Analyse des Correspondances, pas plus que la contribution (Gionis et al., 2007) qui aborde comme nous le problème du nombre de dimensions significatives d'une matrice binaire rectangulaire, mais de façon heuristique, en se basant sur une unique matrice randomisée.

L'approche analyse Sémantique Latente, ou LSA (Deerwester et al., 1990), permet de calculer des similarités angulaires entre unités statistiques textuelles ou entre descripteurs dans un espace de dimensions réduites par décomposition aux valeurs singulières de la matrice des occurrences des mots dans les textes, pondérée « tf-idf ». Cette réduction est différente de celle effectuée par l'AFC, et également à la base du partitionnement spectral de graphe comme montré plus loin, dont elle n'a pas les fondements théoriques. D'autre part le nombre de dimensions de l'espace réduit est l'objet de recommandations empiriques (« 300 à 500 pour quelques milliers de textes ») et non de critères spécifiques.

3. Espace intrinsèque d'une matrice binaire

C'est le détour par les graphes qui nous permettra d'établir notre procédure d'extraction du nombre et de la nature des dimensions intrinsèques d'une matrice binaire. Définissons d'abord la notion d'espace intrinsèque : l'espace intrinsèque d'un graphe (non orienté, non valué) est l'espace de représentation réduit dans lequel se trouvent concentrées et mises en évidence ses caractéristiques structurelles « intéressantes », (regroupements de nœuds en *clusters* ou en chaînes, …). Ces caractéristiques le différencient de ses variantes randomisées, c'est à dire à même répartition des degrés des nœuds, mais à répartition aléatoire des liens.

3.1. AFC et laplacien d'un graphe

A notre connaissance, la première application aux graphes de l'analyse spectrale remonte à (Benzécri, 1973) (qui reprenait le cours polycopié de 1969 *Sur l'analyse de la correspondance définie par un graphe*), dans lequel l'Analyse Factorielle des Correspondances (AFC) était appliquée à la matrice d'adjacence d'un graphe. Rappelons que l'AFC (Greenacre, 2007 ; Lebart et al., 1984) repose sur la décomposition aux valeurs singulières d'une matrice **Q** issue du tableau de correspondance **X** : $\mathbf{Q} = \mathbf{D_r}^{-1/2}\ \mathbf{X}\ \mathbf{D_c}^{-1/2}$ où $\mathbf{D_r}$ et $\mathbf{D_c}$ sont les matrices diagonales des sommes en lignes et en colonnes (à noter que pour un graphe non orienté et non pondéré, on applique une telle décomposition à une matrice d'adjacence **X** symétrique et à valeurs binaires). La décomposition aux valeurs singulières de **Q** s'écrit $\mathbf{Q} = \mathbf{U}\ \mathbf{\Lambda}\ \mathbf{V'}$, où $\mathbf{\Lambda}$ est la matrice diagonale des valeurs singulières (parmi lesquelles : $\lambda_1 ... \lambda_L = 1$, L étant le nombre de composantes connexes ; et $1 > \lambda_{L+1} > ... > \lambda_R > 0$, R étant le rang de **X**). Les matrices **U** et **V** rassemblent les vecteurs propres pour les lignes et les colonnes respectivement, donnant lieu à plusieurs variantes possibles des facteurs, au gré des auteurs.

(Benzécri, 1973) a proposé des solutions analytiques pour des graphes simples comme les anneaux ou les grilles. Dans (Lebart, 1984) l'auteur a généralisé à l'analyse de la contiguïté, et illustré en montrant que le plan factoriel (F2, F3) de l'AFC de la matrice de contiguïté entre les départements français reconstituait l'allure de la carte de France.

Une lignée de recherche indépendante initiée par (Chung, 1997) a défini deux matrices "laplaciens normalisés de graphes", à savoir le laplacien symétrique (**I** – **Q**), où **I** est la matrice identité (on a $\lambda_1,..., \lambda_L = 0$, L étant le nombre de composantes connexes ; $0 < \lambda_{L+1} < ... < \lambda_R$, R étant le rang of **X**. et sa variante "marche aléatoire" $\mathbf{I} - \mathbf{D_r}^{-1}\ \mathbf{X}$ - on note que les valeurs propres de **Q** sont les compléments à 1 de celles de **I-Q.**

La partition spectrale de graphe consiste à grouper les nœuds dans l'espace des K plus importants vecteurs propres – pour une revue cf. (Chung, 1997) – et constitue une voie de recherche de plus en plus active. Jusqu'à présent, à notre connaissance, la détermination du nombre K, quand la distribution des degrés sort des modèles classiques (loi binomiale, etc.),

n'a pas reçu de réponse plus satisfaisante que la classique détermination visuelle ou par examen des différences secondes d'une discontinuité dans la séquence des valeurs propres (Cattel, 1966) – ce qui ne pose pas de problème pour les petits graphes, mais passe difficilement à l'échelle de centaines ou milliers de nœuds. Sans parler de la très heuristique suggestion K = √(Rang de la matrice des données)…

3.2. Dimension intrinsèque : un test de randomisation

Notre but n'est pas de simuler un graphe aléatoire avec une suite de degrés donnés, mais, pour un graphe donné, biparti ou non, de générer de la façon la plus directe et rigoureuse possible une suite de graphes aléatoires indépendants de même suite de degrés. Les deux points de ce cahier des charges ne sont pas remplis par le « configuration model » (Molloy et Reed, 1995) (qui ne garantit pas l'adéquation rigoureuse à une distribution des degrés donnée) et ses dérivés, qui compliquent ce modèle, cf. par exemple (Viger et Latapy, 2005).

TourneBool (Cadot, 2006) est une méthode de génération de N versions aléatoires ("randomisées", N souvent égal à 100 ou 200) d'un tableau de données binaires, à marges lignes et colonnes inchangées, et de test statistique de toute quantité construite sur ce tableau, par comparaison avec les N valeurs trouvées sur les tableaux randomisés. Il est à noter que les principes de génération de matrices aléatoires à marges fixes, en partant d'une matrice binaire donnée, semblent avoir été découverts indépendamment plusieurs fois dans plusieurs domaines d'application : écologie, psychométrie, sociologie, combinatoire. Pour ce qui nous concerne, (Cadot, 2005) a présenté un algorithme de permutation basé sur des échanges rectangulaires (un échange rectangulaire à la croisée des lignes i1 et i2 et des colonnes j1 et j2 est possible sans modifier les marges si les cases (i1,j1) et (i2,j2) valent 1 alors que les cases (i1,j2) et (i2,j1) valent 0) ; il incorpore un contrôle de la convergence de l'algorithme pour éviter tout biais. Sa justification théorique, exposée dans (Cadot, 2006), est basée sur la notion d'échange en cascade, opération qui transforme une matrice booléenne en une autre matrice de mêmes marges – et à l'inverse, il a été montré dans ce même mémoire que toute matrice booléenne pouvait être transformée en toute autre de mêmes sommes marginales en un nombre fini de telles cascades. Dans le domaine des graphes, nous avons appliqué cette approche pour créer des graphes de liens (et d'anti-liens) valides entre variables booléennes (les mots) à partir de corpus textuels (Lelu et Cadot, 2010).

Comme c'est le cas pour tous les autres tests de randomisation (Manly, 1997), l'idée générale vient du test exact de Fisher (Fisher, 1936), mais elle concerne les variables prises comme un tout, et non deux à deux Les échanges élémentaires préservent la structure d'arrière-plan irréductible de la matrice des données, mais brisent les liens chargés de sens qui caractérisent les données issus de la vie réelle. Par exemple, la plupart des matrices textes × mots ont une distribution des mots en loi de puissance, et une distribution d'allure binomiale du nombre de mots uniques dans les textes. Cette structure d'arrière-plan conditionne notre espérance statistique d'absence de lien sachant le type de corpus. La neutraliser permet de traiter tout type de données binaires, à la fois en prenant en compte les distributions marginales , et en le faisant sans avoir à spécifier un modèle statistique pour ces distributions. Les paramètres de l'algorithme sont au nombre de trois : le nombre d'échanges nécessaires pour engendrer des matrices aléatoires non biaisées, le nombre de matrices à créer, le risque alpha.

A noter que les tests de permutation, dont dérivent les tests de randomisation, ont été démontrés comme les plus "puissants", c'est à dire minimisant le risque bêta pour un risque alpha donné (Droesbeke et Finne, 1996).

3.3. Cas d'une matrice binaire quelconque (graphe biparti)

Un résultat bien établi en analyse des données dit que l'information pertinente, débarrassée du bruit, réside dans les éléments propres dominants d'une matrice de données [8]. Dans le cas de la matrice **Q** (Benzécri, 1973 ; Chung 1997) et beaucoup d'autres ont montré que la valeur propre dominante, de multiplicité L (L étant le nombre de composantes connexes du graphe) est 1 – il en va de même pour la matrice $\mathbf{D_r}^{-1}\mathbf{X}$.

Dans le cas d'un graphe biparti, dont la matrice d'adjacence et le Laplacien symétrique s'écrivent respectivement $\begin{vmatrix} \mathbf{0} & \mathbf{M} \\ \mathbf{M'} & \mathbf{0} \end{vmatrix}$ et $\begin{vmatrix} \mathbf{0} & \mathbf{Q} \\ \mathbf{Q'} & \mathbf{0} \end{vmatrix}$, une simplification découle de la propriété qu'ont ces matrices d'avoir leurs valeurs propres composées 1) des valeurs singulières de leur sous-matrice rectangulaire non-vide, 2) des opposées de ces valeurs – dans le cas de **Q**, dans l'intervalle [-1:+1]. Il s'ensuit que c'est au « tunnel de confiance » situé entre la limite supérieure de, par exemple, 95% des valeurs propres positives *et leurs opposées* qu'il faut comparer la séquence des valeurs propres de la matrice d'origine. La figure 1 illustre cette situation dans le cas d'un graphe-jouet de 66 nœuds organisés en 4 *clusters* bruités, au sein d'une même composante connexe.

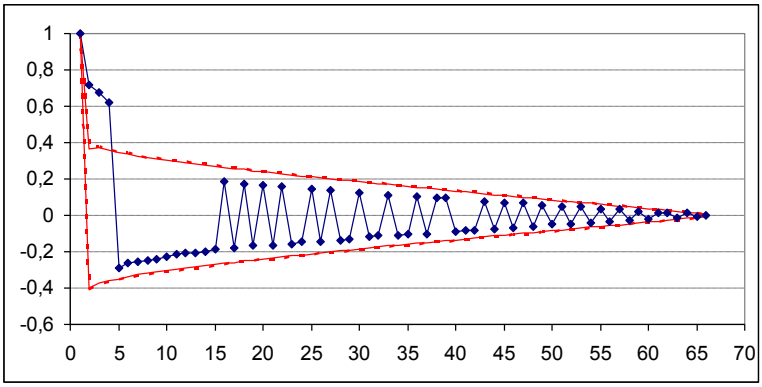

Figure 1. Graphe avec 4 clusters bruités, distincts dans les 3 premières dimensions propres (hors de la dimension 1, triviale) : tunnel de confiance (rouge) des valeurs propres (bleu).

Dans le cas de l'AFC l'opération de base, c'est à dire la décomposition aux valeurs singulières de son laplacien symétrique **Q**, constitue la référence pour comparer cette matrice à ses homologues aléatoires, et il suffit d'observer le rang de la dernière valeur singulière, en partant de $\lambda_1=1$ située au-dessus de l'enveloppe des valeurs singulières des matrices aléatoires générées.

4. Les données et leur pré-traitement

Dans le cadre de la veille sur les controverses Sciences-Société assurée par le CNRS/ISCC, 789 articles de presse complets ont été tirés de la base Lexis-Nexis au sujet de la controverse OGM et de celle des perturbateurs endocriniens (Bisphénol A, …) au 1er trimestre 2011, dans la presse tant régionale que nationale. Pour le choix des descripteurs des textes, il était souhaitable à la fois de maximiser la précision sémantique, donc de minimiser les ambiguïtés, et de réduire la taille de l'espace de description, c'est à dire le nombre de descripteurs. Notre choix s'est porté sur les couples lemme + étiquette syntaxique. En effet l'extraction de termes composés, généralement non ambigus, accompagnés de lemmes possiblement ambigus, utilisée jusqu'à présent dans nos travaux, aurait agrandi à l'excès l'espace de représentation

(près d'une douzaine de milliers d'éléments, pour 789 textes), ceci s'ajoutant au bruit dû aux lemmes ambigus. A noter que nous avons tenu à n'éliminer aucun lemme, fût-il grammatical ou autre, éprouvant en cela la robustesse de notre approche aux répartitions très inégalitaires, zipfiennes, des mots et des ponctuations, qui peuvent différer pour certains de deux ou trois ordres de grandeur.

4.1. Lemmatisation et étiquetage morpho-syntaxique : l'étiqueteur Morph

MORPH, dérivé batch de l'étiqueteur et analyseur syntaxique LLP2 a été développé par Azim Roussanaly <*www.loria.fr/~azim/LLP2/help/fr*>. A la différence de *TreeTagger* qui utilise un lexique construit à la volée sur un corpus d'apprentissage, il utilise la ressource externe disponible la plus complète à notre connaissance pour le français, le dictionnaire Morphalou (~540 000 formes, ~95 000 lemmes) du CNRTL, <www.cnrtl.fr/lexiques/morphalou/>. L'apprentissage se fait de manière « condensée » par HMM (modèles de Markov cachés) sur les seules séquences d'étiquettes du corpus d'entrainement. Il utilise le jeu d'étiquettes de l'ATILF, condensé ici à 25 codes (SBC:nom commun, ADJ:adjectif, ADV:adverbe, …) + 3 codes-préfixes de verbes (A:avoir, E:être, V:autres verbes) + la ponctuation en clair. Les mots inconnus ont reçu l'étiquette U[*nknown*], assimilée par défaut à SBC. Le programme (en Java) met en œuvre une recherche rapide dans le dictionnaire par arbre ternaire, qui le rend efficace : ici 3' pour traiter les 2,5 Mo du corpus sur un PC *Quadcore Intel* 2,6 MHz. Sa précision mesurée par validation croisée sur le corpus d'apprentissage atteint 98%.

Limites : les défauts de ponctuation (parfois inexistence du point entre deux phrases !) et les « non-phrases » du corpus (titres et méta-informations hétérogènes) peuvent conduire parfois à lemmatiser « venaient » par le verbe ancien « vener », « [qu'ils] aillent » par « aller », « Agence [France-Presse] », par « agencer », … Idéalement, une détection d'entités nommées (ou de candidats entités nommées) serait à insérer en début de chaîne, au prix toutefois d'un important travail humain de validation.

4.2. Les sorties possibles

Le fichier de sortie comporte pour chaque forme, son couple lemme-étiquette le plus probable, compte tenu des étiquettes gauche et droite. De cette sortie ont été dérivés deux fichiers : une liste « en clair » des 26 942 lemmes étiquetés du corpus (hapax compris), et la séquence des 461 436 numéros de lemmes étiquetés qui le constituent. Cette séquence est disponible pour découper le corpus en unités statistiques de granularité différente, selon le choix des séparateurs, par exemple : §§§§ (séparateur d'article) pour un découpage en articles, ou . ; : ? ! ... §§§§ pour un découpage en phrases. Il en résulte des matrices [unités statistiques × descripteurs], dans notre cas une matrice **X1** de 789 articles × 7 499 lemmes de fréquence >3, et une matrice **X2** de 32 917 phrases (de plus d'un mot) × 7 081 lemmes de fréquence >3. Ces matrices sont alors rendues binaires : 1 pour toute présence du lemme i dans le document j, 0 sinon. D'autres séparateurs seraient nécessaires pour un découpage en virgulots, et la séquence des n° de lemmes permettrait un découpage en N-grammes.

5. Les traitements

On a choisi deux découpages contrastés du corpus : celui par articles (789 textes), et celui par phrases (32 917 phrases de plus d'un mot). Nous laisserons de côté dans la présente étude les découpages plus fins (en virgulots, en N-grammes).

5.1. Détermination et exploitation de l'espace intrinsèque des mots dans le contexte des articles

5.1.1. Test de randomisation

On crée par TourneBool 200 matrices (789×7 499), dérivées randomisées de la matrice des observations **X1**, dont on calcule les valeurs singulières.

Résultat : au seuil de significativité de 1%, les valeurs singulières de **X1** pénètrent l'intervalle de variation aléatoire entre les valeurs n°196 et 197 (cf. figure 2). La dimension intrinsèque est donc de 196-1 (la 1ère valeur propre est triviale), i.e.195.

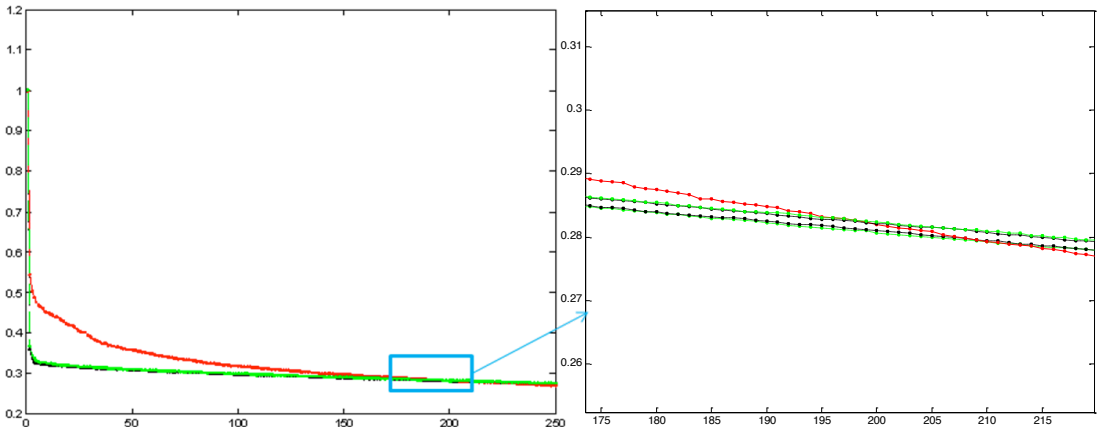

*Figure 2. L'éboulis des 250 premières valeurs singulières : celles de **X1** (en rouge) pénètrent l'intervalle de variation aléatoire au seuil de 1% (en vert) entre les valeurs n°196 et 197.*

5.1.2. Liens et anti-liens entre mots dans ces 195 dimensions

En appliquant la méthode décrite dans (Lelu et Cadot, 2010) au sein des distances entre mots dans l'espace des 195 premiers vecteurs singuliers (matrice « V195 » : 7 499 lignes, 195 colonnes), on obtient le graphe de leurs liens statistiquement valides on compare à chaque valeur de la matrice des cooccurrences (**X1' X1**) la série des 200 valeurs correspondantes créées par les variantes randomisées, qu'on range par ordre croissant : si la 199ème est inférieure à cette valeur et différente de zéro, le lien est réputé valide au seuil de significativité de 1%. ; de même pour les anti-liens. Un sondage rapide sur les résultats montre que ces liens sont cohérents : par exemple le mot *image* est lié à *choc, France, affichage, affiche, campagne, algue, vert, nature, ne, pas, provocation, salarié*… ; et antilié à *proposition, membre, Commission, collectif, génétique, futur, revue, loi, Espagne, autorisation, contrôle*…Cette dernière liste, plus difficile à interpréter, comporte beaucoup de mots du vocabulaire juridique et institutionnel, contexte dans lequel il est compréhensible qu'*image* apparaisse significativement moins qu'ailleurs. Il reste à explorer l'utilisation de ces filtrages en recherche d'information, en comparant leur utilisation aux performances obtenues sur des corpus de test d'accès public.

5.1.3. Partitions des mots par la méthode des k-moyennes axiales dans ces dimensions

A une partition par la méthode des *k-means*, comme pratiquée habituellement en « *spectral graph clustering* », nous avons préféré celle des *k-means* axiales, décrite en détail et revisitée dans (Lelu et al., 2013), qui présente l'avantage de fournir des listes d'éléments classés par ordre décroissant de centralité dans chaque *cluster*, entre autres avantages (nous n'avons pas

exploité ici celui permettant de rattacher un même mot à plusieurs contextes, qui peut rendre compte d'effets de polysémie).

En demandant 195 *clusters* de mots dans cet espace à 195 dimensions, nous avons obtenu une majorité de *clusters* de taille moyenne, autour de 100 mots. Un seul de ces *clusters* rassemble des éléments syntaxiques et de ponctuation : *que* | *pour* | *il* | *qui* | *pas* | . | *à+le* | *§§§§* | *:* | *le* | … A l'exception du plus gros *cluster*, de taille 400 environ et de caractère « fourre-tout » semble-t-il, les autres sont centrés sur des « histoires » qui reviennent de façon récurrente dans plusieurs titres de presse ou dépêches d'agence.

Ex. d'« histoire » : *le commissaire européen Dacian Ciolos discute agriculture à Washington* revient dans 4 articles, dont 2 redondants, et donne lieu au *cluster* de mots suivant, de haut en bas et de gauche à droite :

antimicrobiens_U	Dacian_U	durant_PREP	répondre_PAR
spongiforme_U	CIOLOS_U	494_CAR	conserver_VNCFF
Peterson_U	impression_SBC	rappeler_PAR	subvention_SBC
Ron_U	09-févr_SBC	préparation_SBC	
Mike_U	Lucas_U	US_SBC	

On voit donc que l'espace intrinsèque traduit principalement des liens sémantiques à moyenne portée entre mots, réunis dans une grosse centaine de narrations qui ont fait l'actualité journalistique du 1er trimestre de 2011, dans un large domaine centré sur les controverses « OGM » et « perturbateurs endocriniens ».

5.2. Détermination et exploitation de l'espace intrinsèque des mots dans le contexte des phrases

On crée par TourneBool 200 matrices (32 917×7 081), dérivées randomisées de la matrice des observations **X2**, dont on calcule les valeurs singulières : au seuil de significativité de 1%, les valeurs singulières de **X2** pénètrent l'intervalle de variation aléatoire entre les valeurs n°2 744 et 2 745. La dimension intrinsèque est donc de 2 744-1 (la 1ère valeur propre est triviale), i.e. *2 743*. Ce nombre est considérable, et c'est déjà un résultat important en regard de la règle empirique des « 400 premières dimensions » préconisées en Analyse Sémantique Latente (LSA) pour des corpus textuels.

5.2.1. Exploration des partitions de mots par la méthode des k-moyennes axiales dans tout ou partie de ces dimensions

Ce nombre pourrait donner lieu, en théorie, à près de 3 000 *clusters* pour les quelques 7 000 mots à partitionner, dont une certaine proportion de *clusters* constitués d'un mot isolé[1]. Deux difficultés se présentent :

- Une difficulté technique due aux problèmes d'espace mémoire et d'initialisation du programme de *clustering* si on lui demande de créer un nombre de catégories (~3 000) du même ordre de grandeur que les éléments à répartir (~7 000).

[1] Le test TourneBool a été appliqué à la validité des liens entre mots, comme décrit en 7.1.2, puis l'algorithme classique de parcours en profondeur a permis de détecter les composantes connexes : une composante géante de 6 490 mots, 3 composantes de 3 mots (ex : Christiane | Lambert | Bouchart), 9 de 2 mots (Serge | Dassault ; malade | SIDA ; laïque | obscurantisme ; …), et 578 mots isolés.

- Une difficulté ergonomique : si l'utilisation d'un espace à 2 700 dimensions ne présente pas de problème pour effectuer une tâche supervisée et l'évaluer par un indicateur numérique comme la précision ou la F-mesure, il n'en va pas de même dans un cadre non-supervisé, où il va de soi que l'esprit humain a de la difficulté à appréhender plus de quelques dizaines de catégories simultanément, et à effectuer sur celles-ci un travail sérieux de comparaison. C'est pourquoi nous aborderons l'exploitation d'un tel espace intrinsèque de façon progressive et exploratoire. Comme les dimensions sont rangées par ordre décroissant d'importance nous nous intéresserons d'abord aux 5 premières, puis aux 50 premières, puis aux 500 premières, et enfin à la totalité (matrice « V2743 » : 7 081 lignes, 2 743 colonnes).

5.2.2. Partitions dans V5

La partition en 5 *clusters* demandée conduit de manière remarquablement stable à deux *clusters* importants (autour de 5 300 et 1 000 mots respectivement), deux petits *clusters* (autour de 90 mots) et un *cluster* moyen (autour de 600 mots). L'interprétation des petits *clusters* ne pose pas de problèmes : l'un contient les noms des journaux et agences de presse (généralement en majuscules) et comme seuls verbes *grouper, presser* et *agencer*, issus d'analyses erronées des zones de métadonnées concernées… L'autre contient les mots des titres de rubriques (*TELEVISION, REGIONS, OPINIONS,* …) et le seul verbe *coter* issu d'une erreur d'analyse de la rubrique *COTE*. Notre traitement filtre donc correctement les métadonnées, dont le format ne présente aucune homogénéité dans le corpus, et seraient très difficiles à extraire de façon rigoureuse. L'interprétation des autres *clusters* est plus délicate, et nous avons fait appel à des comptages sur les 100 premières étiquettes syntaxiques pour tenter d'y voir plus clair :

- Le *cluster* moyen contient peu de verbes et de participes (moins de 10%) et beaucoup de termes non reconnus (20%), qui suggèrent, avec les noms propres et communs, le thème général « écologie et politique ».

- Le *cluster* majoritaire (75% du vocabulaire) comporte la plus grande proportion de verbes et participes (un mot sur 3) dans sa tête de liste, et remarquablement aucun mot non reconnu. Son vocabulaire dégage une tonalité plutôt économique et institutionnelle (*'16 : régie_SBC' '9 : brut_ADJ' '11 : automobile_ADJ' '9 : exportateur_ADJ' '14 : crainte_SBC' '10 : OMC_SBP' '8 : nettement_ADV' '8 : organisateur_SBC' '13 : Isabelle_SBP' '18 : siège_SBC' '8 : 250_CAR' '9 : betterave_SBC' '8 : indication_SBC' '10 : écart_SBC' '8 : offshore_SBC' ...*) – les nombres qui précèdent les mots sont leurs occurrences. La liste des verbes (*'10 : bénéficier_VNCFF' '8 : rapporter_VCJ' '8 : conclure_VNCFF' '9 : entraîner_VNCFF' '8 : rassembler_VNCFF' '7 : avouer_VCJ' '12 : contrôler_VNCFF' '12 : parvenir_VNCFF' '8 : receler_VCJ' '7 : affecter_VCJ' '20 : jouer_VNCNT' '7 : réaliser_VCJ' '7 : citer_VNCFF' ...*) confirme apriori ce ton de narration neutre d'événements de l'actualité – à coup sûr la base de l'écriture journalistique.

Le deuxième plus gros *cluster* présente une moindre proportion de verbes et participes (un mot sur 5), mais fait apparaître environ 10% de mots inconnus, et autant de nombres (codés – CAR). Il est difficile de dégager une thématique à partir des noms, mais la liste des verbes (*'5 : subir_VCJ' '5 : planter_VNCFF' '4 : balayer_PAR' '4 : opposer_VNCNT' '5 : ordonner_VCJ' '5 : pencher_VCJ' '8 : diriger_VCJ' '5 : apparaître_VNCFF' '4 : remonter_VCJ' '7 : régir_VCJ' '7 : fixer_VNCNT' '7 : attirer_VCJ'*) suggère la narration d'actions militantes, peut-être le signe d'un récit plus engagé, ou des phrases transcrivant les propos des acteurs. Ce qui serait cohérent avec une proportion plus grande de néologismes inconnus du dictionnaire du CNRTL.

5.2.3. Partitions dans V50

La répartition des effectifs de mots dans les 50 *clusters* demandés varie beaucoup plus que précédemment avec la graine d'initialisation. Sur 20 essais, les valeurs minimales et maximales du pourcentage de reconstitution des données ont été de 56,92% et 59,28% respectivement. C'est ce dernier passage qu'on a retenu : 19 *clusters* rassemblent moins de 10 mots, 18 en ont de 10 à 63, 11 en ont de 108 à 711, et deux dépassent les 1 000 mots :

- Les petits *clusters* se spécialisent sur certains micro-aspects des métadonnées, par ex. *{&* | *SANTE* | *FORME* | *VOUS* | *OPINIONS}* ou *{Express* | *Expansion* | *groupe* | *grouper* | *magazine* | *-Online* | *Echos* | *Marianne}*, *{CONTRE* | *-ENQUETE}*.

- Le plus gros *cluster* (1 832 mots) semble consacré à la description des problèmes douloureux de la filière agricole.

- Le deuxième plus gros *cluster* (1 254 mots) semble spécifique des aspects juridiques, institutionnels et diplomatiques des crises environnementales, avec le style caractéristique du journalisme « respectable » dont l'idéal-type est le style soutenu du journal Le Monde.

- Les *clusters* de taille intermédiaire semblent rendre compte d'aspects transversaux : environnement et politique régionale / nationale, la malbouffe, les procès des faucheurs volontaires, le mouvement écobio, …

5.2.4. Partitions dans V500 et dans V2743

Le compromis stabilité/exhaustivité inclus dans les choix d'initialisation n'a pas permis de créer plus de 350 *clusters*, dont la moitié comporte moins de 10 mots. Les regroupements sont plus fins : noms d'hommes politiques, de journaux, thèmes sans verbes (transgénie, intoxications), procès Berlusconi, vocabulaire du compromis et des suggestions positives, …

5.2.5. Conclusion sur le découpage en phrases

Le découpage d'un corpus en phrases et le *clustering* basé sur la cooccurrence des mots dans ces phrases, après réduction de l'espace des données, dégagent des éléments de style : style militant, style de journalisme « respectable », style de compte rendu judiciaire, … Très globaux quand on découpe un petit nombre de *clusters* dans un petit nombre de dimensions, ils deviennent de plus en plus spécifiques et liés à un contenu sémantique au fur et à mesure qu'on augmente la finesse d'analyse (nombre de *clusters* et nombre de dimensions). On se heurte alors à l'instabilité constitutive des méthodes de *clustering* de type *k-means*, et d'autres solutions seraient à envisager, comme les méthodes par densité.

6. Conclusions, perspectives

Nous avons exploré ici la possibilité, généralement exclue dans les analyses de textes, de prendre en compte simultanément des descripteurs d'occurrences différant de 4 ou 5 ordres de grandeur, fréquents dans les répartitions *zipfiennes*, depuis les déterminants et signes de ponctuation courants jusqu'aux mots de fréquence 4. Ceci pour éviter des filtrages souvent arbitraires, pour mettre au jour des relations intéressantes entre éléments de différents niveaux dans le discours ; mais ceci nécessite une exploitation multi-échelle, que nous avons testée de deux façons différentes : 1) d'abord en variant les tailles des unités textuelles élémentaires considérées : le découpage en articles a fait apparaître prioritairement des regroupements thématiques par « histoire » ou sujet commun traité par plusieurs articles (ou issus de la même source ?). Le découpage par phrases a fait apparaître des éléments communs de style d'écriture et d'expression composées, mais souvent mélangés à des éléments thématiques.

Pour éviter cela un découpage par virgulots serait préférable. Le découpage le plus fin, par N-grammes de mots, est déjà connu comme base pour la tâche d'étiquetage morphosyntaxique non supervisée d'un corpus de langue quelconque ou inconnue. 2) ensuite, via le principe de la partition spectrale de graphes : les K premières dimensions de l'espace propre réduit extrait du laplacien du graphe inter-mots forment l'espace-support d'une partition en K *clusters* au moyen d'une méthode de type *k-means*. Pour les K au-delà de 350 nous nous sommes heurtés aux limites de ces méthodes en termes d'initialisation et d'optima locaux, ainsi qu'aux limites ergonomiques quand il s'agit d'examiner et interpréter des centaines de *clusters*. Mais les valeurs intermédiaires donnent des *clusters* interprétables en termes de style d'écriture journalistique et/ou de domaine couvert (actions militantes, réglementation, procès…), où les métadonnées (titres de presse, libellés de rubriques, …) se détachent toujours en dépit de leur faible normalisation. Les matrices de liens et d'anti-liens entre mots peuvent aussi donner directement lieu à des exploitations fines pour qui s'intéresse à tel ou tel mot de la rhétorique journalistique.

Sur un plan plus théorique, nous avons montré la validité et l'extension de l'approche Analyse des Correspondances (AFC) dans ses liens avec la notion de laplacien normalisé d'un graphe non orienté. Pour aller plus loin en pratique, la prise en considération de seuils de significativité plus drastiques serait à considérer en premier lieu pour multiplier les composantes connexes, puis il faudra envisager de passer à d'autres méthodes de *clustering*, plus stables et fines que celles de type centres mobiles, comme celles de densité ; enfin il faudra concevoir des aides automatisées (graphiques ?) au dépouillement et à l'interprétation d'un grand nombre de *clusters*, afin de porter la justification subjective du nombre de dimensions intrinsèques au même niveau que sa justification par des critères numériques, en apprentissage supervisé, que nous avons publiée dans (Lelu et Cadot, 2011). A titre d'exemple, l'application dans le cadre d'analyse de bases de textes ou résumés scientifiques devrait permettre de distinguer (« typer ») les mots relevant d'outils transversaux à plusieurs disciplines, en regard de ceux relevant des thématiques disciplinaires traitées.

Remerciements

A l'Institut des Sciences de la Communication du CNRS (ISCC) pour nous avoir fait bénéficier de son travail de veille et de filtrage d'information réalisé sur la base de presse Lexis-Nexis. A Michel Zitt pour avoir attiré notre attention sur l'attente de typage des mots en recherche d'information et scientométrie.

Références

Benzécri J.-P. (1973). *L'analyse des données* (3 tomes) Dunod.

Bouveyron C., Celeux G. et Girard S. (2007). Intrinsic Dimension Estimation by Maximum Likelihood in *Probabilistic PCA, Statistics and Computing* 17(4).

Cadot M. (2005). A simulation technique for extracting robust association rules. In: CSDA.

Cadot M. (2006). Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association. PhD thesis, Franche-Comté University.

Cattell R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2),. 245-276.

Chung F.R.K. (1997). Spectral Graph Theory, (*CBMS Regional Conference Series in Mathematics*, No. 92), American Mathematical Society.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K. et Harshman R. (1990). Indexing by Latent Semantic Analysis. *JASIS* 41 (6): 391–407.

Droesbeke J.J. et Finne J. (1996) Inférence non paramétrique. *Les statistiques de rangs*. ASU & Ellipse.

Fisher R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, pp 179–188.

Gionis, A., Mannila, H., Mielikäinen, T. et Tsaparas, P. (2007). Assessing data mining results via swap randomization. *ACM Trans. Knowl. Discov. Data*.

Girvan M. et Newman M. E. J. (2002). Community structure in social and biological networks, *Proc. Natl. Acad. Sci. USA* 99, 7821–7826

Greenacre M. (2007). *Correspondence Analysis In Practice* (interdisciplinary Statistics) Chapman & Hall/crc Interdisciplinary Statistics Series.

Lancichinetti A. et Fortunato S. (2009). Benchmark for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review*. E 80.

Lebart L., Morineau A. et Warwick K. (1984). *Multivariate Descriptive Statistical Analysis*, John Wiley & sons, NY,

Lebart L. (1984). Correspondence Analysis of Graph Structure - In. Comm. Meeting of the Psychometric Society , *Bulletin Technique du CESIA*, vol 2, 5-19.

Lelu A. et Cadot M. (2011), Espace intrinsèque d'un graphe et recherche de communautés. *Revue I3*, vol.11, pp.1:25, CEPADUES, Toulouse.

Lelu A. (2010). Slimming down a high-dimensional binary datatable: relevant eigen-subspace and substantial content. In *COMPSTAT 2010*.

Lelu A. et Cadot M. (2010). Statistically valid links and anti-links between words and between documents: applying TourneBool randomization test to a Reuters collection. *Advances in Knowledge Discovery and Management* (AKDM) 292, 307-324.

Lelu A., Zitt M. et Bassecoulard E. (2013). Robustesse des classements bibliométriques, à travers la convergence des thèmes obtenus par citations et lexiques : une méthode hybride pour une représentation mixte – in *Proc. of VSST 2013*, Nancy, 23-25/10/2013.

Manly B. (1997). *Randomization, Bootstrap and Monte Carlo methods*. Chapman and Hall/CRC

Molloy M. et Reed B. (1995). A critical point for random graphs with a given degree sequence, *Random Structures and Algorithms*, 6, 161-179.

Roussanaly A. Morph POS-Tagger: *www.loria.fr/~azim/LLP2/help/fr*; accessed on 02/28/2014.

Schuetze H. (1995). Distributional Part-of-Speech Tagging. *Online Proc. ACL SIDGAT Workshop*.

Viger F. et Latapy M. (2005). Efficient and Simple Generation of Random Simple Connected Graphs with Prescribed Degree Sequence. *COCOON 2005*: 440-449.

Von Luxburg L. (2007). A Tutorial on Spectral Clustering. *Statistics and Computing* 17(4).