La cooccurrence asymétrique : propriétés quantitatives en disponibilité et énergie

Julien Bonneau[1]

[1]Laboratoire BCL – Université Nice Sophia-Antipolis – CNRS UMR 6039 – MSH de Nice – 24 av des diables bleus – 06357 NICE CEDEX 4 – France

Abstract

As for a target word of a text, we can define two relationships: the part given to this target word by every word it is related to in a co-text; the part given by this target word to every word it is related to in the co-text. So, that defines two assymetric co-occurrences. Are these two indexes of word co-textualisation in a text, relevant to describe its properties? Are these indexes distorted by the size of the texts which are studied? These are the question we are raising to a corpus of various textual genres by the same author, Pierre Mendès France, in order to open pathways to an automatic text comparison method.

Résumé

Pour un mot donné, on peut définir la relation cooccurrentielle comme la part que prennent chacun des mots qu'il rencontre en cotexte ou comme la part que donne ce mot pôle à chacun des mots qu'il rencontre en cotexte. On définit ainsi deux relations asymétriques de cooccurrence, l'une dite en énergie, l'autre en disponibilité. Ces deux indices de cotextualisation des occurrences d'un texte sont-ils pertinents pour décrire ses propriétés ? Ces indices sont-ils soumis aux effets de taille des textes étudiés ? Ce sont les questions que nous soumettons à un corpus de textes de Pierre Mendès France, variés en genre, dans le but d'ouvrir des pistes vers une méthode automatique de comparaison de textes.

Keywords : asymmetric co-occurrence, scatter plot, threshold effect, R Project, political corpus.

1. Introduction

Pour un mot donné, on peut définir la relation cooccurrentielle comme la part que prennent chacun des mots qu'il rencontre en cotexte ou comme la part que donne ce mot pôle à chacun des mots qu'il rencontre en cotexte. On définit ainsi deux relations asymétriques de cooccurrence, l'une dite en énergie, l'autre en disponibilité (Luong *et al.*, 2010) que l'on peut représenter dans une même matrice dite matrice énergie-disponibilité.

Les distances intertextuelles classiques (Brunet, 2003), type Jaccard ou Labbé, ne sont pas adaptées au traitement des matrices énergie-disponibilité des textes, celles-ci gommant les relations entretenues par cette double composante ou les assimilant à un même résultat. Notre méthode consiste à représenter dans un plan tous les lemmes d'un texte, l'énergie constituant l'abscisse, la disponibilité l'ordonnée. On compare ensuite visuellement les nuages-textes

constitués. Une première étude (point 3) s'attachera à décrire les effets de taille sur un corpus stable en genre, une deuxième (point 4), à comparer les profils de textes de genre différent, mais de taille semblable. La première analyse proposée ne s'intéresse qu'à la géométrie des nuages énergie-disponibilité dans le plan ; la seconde projette sur chacun des nuages constitués, la qualité morphosyntaxique des lemmes points, espérant ainsi faire apparaître, d'un genre textuel à l'autre, des dispositions spatiales spécifiques des parties du discours.

2. Les données et leur traitement

2.1. Calcul et représentation des cooccurrences

On entend par cooccurrence, la coprésence de deux mots dans une même unité cotextuelle. Ces unités peuvent être définies comme un nombre de mots autour d'un mot pôle ou des fenêtres fixes de par le texte, par exemple les phrases ou, dans notre cas, les paragraphes. L'information retenue par une unité cotextuelle n'est pas stable, elle dépend non seulement du texte, mais aussi des formes du texte étudié (Martinez : 141). Pour bien faire, la justification d'un tel choix mériterait de nombreuses analyses de comparaison des unités cotextuelles (Martinez, 2003 : 132-142). Dans notre cas, nous faisons simplement l'hypothèse que le paragraphe apparaît comme un réseau sémantique de mots assez proches pour garantir les résonances isotopiques entre les mots qui le composent, mais assez vaste pour espérer une forme de « complétude » du sens. Ce serait donc l'approximation d'une unité méso-sémantique, l'encadrement[1] d'une unité inter-sémique liant les mots qui le compose.

Du point de vue du décompte des cooccurrences, et pour simplifier le calcul, qu'un mot n'apparaisse qu'une ou plusieurs fois dans un paragraphe, il ne sera dénombré qu'une occurrence dans la fenêtre cotextuelle.

Notre étude propose une approche généralisée (Viprey, 1997) de la cooccurrence asymétrique en énergie et en disponibilité sur l'ensemble des lemmes d'un texte. Soient les formes du corpus m_1 , m_2 , …, m_n et A un texte du corpus. Soit a_{ij} la cooccurrence des formes m_i et m_j dans le texte A. On a donc $a_{ij} = a_{ji}$. La matrice énergie-disponibilité de A est donnée par :

$$A = \begin{pmatrix} \frac{a_{11}}{a_{11}} & \cdots & \frac{a_{1n}}{a_{11}} \\ \vdots & \ddots & \vdots \\ \frac{a_{1n}}{a_{nn}} & \cdots & \frac{a_{nn}}{a_{nn}} \end{pmatrix}$$

L'énergie (ou point énergie) de m_j dans A :

$$E_A(m_j) = \sqrt{\sum i \neq j (\frac{a_{ij}}{a_{jj}})^2} = \frac{\sqrt{\sum i \neq j a_{ij}^2}}{a_{jj}}$$

1 Au sens de M. Charolles. Nous voyons le paragraphe comme une relation d'indexation constituant un cadre organisationnel (Charolles, 1997 : 26) produisant de fait une relation sémantique particulière entre ses termes. Ce qui ne dit rien sur la portée de cette relation …

La disponibilité (ou point disponibilité) de m_j dans A :

$$D_A(m_j) = \sqrt{\sum i \neq j \left(\frac{a_{ij}}{a_{ii}}\right)^2}$$

Pour limiter l'effet de la taille du texte[2] sur ces indices, on définit l'énergie et la disponibilité moyenne de m_j, comprises entre 0 et 1. Soit N_A le nombre de formes distinctes présentes dans le texte A (c'est-à-dire le nombre de lignes – ou de colonnes – non nulles de la matrice du texte A) :

L'énergie moyenne (ou point énergie moyen) de m_j dans A :

$$\overline{E_A}(m_j) = \frac{E_A(m_j)}{\sqrt{N_A - 1}}$$

La disponibilité moyenne (ou point disponibilité moyen) de m_j dans A :

$$\overline{D_A}(m_j) = \frac{D_A(m_j)}{\sqrt{N_A - 1}}$$

Pour nous aider à observer et étudier la structure des nuages de points $(E_A(m_j),D_A(m_j))$ et $\left(\overline{E_A}(m_j), \overline{D_A}(m_j)\right)$, on va s'appuyer sur les méthodes d'ajustement linéaire afin de les approcher.

On va détailler dans la suite cette construction pour le nuage $(E_A(m_j),D_A(m_j))$. Il faut pour cela définir :

La moyenne énergétique de A : $m_E(A) = \frac{1}{N_A} \sum_j E_A(m_j)$

La variance énergétique de A : $V_E(A) = \frac{1}{N_A} \times \sum_j E_A(m_j)^2 - m_E(A)^2 \geq 0$

L'écart type énergétique de A : $\sigma_E(A) = \sqrt{V_E(A)} \geq 0$

Ces notions peuvent être définies de façon identique pour la disponibilité du texte A. Pour s'intéresser au nuage $(E_A(m_j),D_A(m_j))$, il faut construire le lien entre la variance énergétique de A et sa variance en disponibilité, la covariance énergie-disponibilité de A :

$$cov_{E,D}(A) = \frac{1}{N_A} \times \sum_j E_A(m_j) \times D_A(m_j) - m_E(A) m_D(A)$$

Nous définirons la droite de régression disponibilité-énergie de A :

$y = ax + b$ où $a = \frac{cov_{E,D}(A)}{V_E(A)}$ et $b = m_D(A) - a \times m_E(A)$

Et la droite de régression énergie-disponibilité de A :

$x = ay + \beta$ où $a = \frac{cov_{E,D}(A)}{V_D(A)}$ et $\beta = m_E(A) - a \times m_D(A)$

Dans nos analyses, le calcul de la matrice énergie-disponibilité a été réalisé à l'aide de script *Perl*. Une fois cette matrice construite pour un texte donné, nous en avons extrait un tableau de contenu suivant : lemme m_i; catégorie grammaticale de m_i ; énergie de m_i ; disponibilité de m_i ; énergie moyenne de m_i ; disponibilité moyenne de m_i. Nous avons ensuite traité ce

2 Ou, plus exactement, l'effet du nombre de mots distincts présents dans le texte.

tableau à l'aide du logiciel *R*[3]. *R* est un logiciel libre permettant de réaliser, ou de programmer, de nombreux traitements statistiques. Il se compose d'un « noyau », cœur du programme permettant d'exécuter des scripts pour réaliser des traitements statistiques, et de « packages » qui sont autant de bibliothèques de scripts (statistiques et/ou graphiques) à disposition. Nous nous sommes appuyés sur le package « *R commander, rcmdr* » (Fox, 2004), qui est une interface graphique conviviale, dont la sélection de tests statistiques paraît plutôt à l'usage des biologistes et des statistiques inférentielles, mais qui permet de charger et d'éditer facilement des données tabulaires (ce qui n'est pas trivial dans *R*) et qui propose, notamment, les représentations en nuage de points[4] et les régressions linéaires dans ses traitements[5].

2.2. Le corpus et les sous-corpus

Les différents textes qui composent notre corpus de travail sont issus de l'ouvrage *Pierre Mendès-France. Œuvres Complètes* (Mendès France, 1984)[6], sélectionnés entre le 18 juin et le 31 décembre 1954, période où P. Mendès France est Président du Conseil des ministres français.

Texte	Nombre de mots	Nombre de phrases	Nombre de paragraphes	Texte	Nombre de mots	Nombre de phrases	Nombre de paragraphes
N15	105	4	2	N20	1044	52	19
IR6	232	11	5	T10	1048	39	21
N23	262	13	6	IR20	1203	41	24
T6	270	11	6	IP7	1464	41	26
IP16	319	10	7	T16	1693	61	22
N25	534	21	12	IP5	2544	77	44
IR7	624	26	12	IP12	5066	158	78
T7	651	16	11	IP20	12047	369	255
IP14	653	21	12	IP11	16217	505	279
IR13	832	33	14				

Tableau 1 : Répartition de formes dans les textes du corpus triés par taille croissante

3 Sous-projet du projet *GNU project* (http://www.gnu.org/gnu/gnu.html) visant à mettre des logiciels gratuits à disposition des utilisateurs, le logiciel *R* est distribué sous *GNU Genaral Public Licence*. Il est développé bénévolement et est distribué *via* des sites *CRAN* (*Comprehensive R archive Network*), dont le principal, http://www.r-project.org/, et de nombreux sites miroirs (en France : le *CICT* à Toulouse ; *IBCP* et *UMR CNRS 5558* à Lyon). Des distributions de *R* existent pour les systèmes d'exploitation *Linux*, *Mac OS X* et *Windows*.

4 Pour améliorer leur définition et aider à leur lecture, nous présentons l'ensemble des graphiques utilisés dans cet article au format pleine page sur le site : http://www.unice.fr/bcl/IMG/pdf/Graphes-Bonneau-jadt2012.pdf

5 Il est à noter que l'on pourra déterminer la référence – le mot – associée à un point par simple clic sur celui-ci. De plus, l'ergonomie du programme est telle qu'il est aisé de reprendre la main sur les traitements effectués. En effet, après l'appel d'une fonction, si sa sortie ne satisfait pas pleinement l'utilisateur, le package *rcmdr* donne automatiquement accès à l'intégralité du script sous-jacent effectué. Il est donc très facile de modifier, détourner, un traitement partiellement satisfaisant pour obtenir le résultat souhaité. De plus, pour toutes les fonctions de *R*, la commande « ?nom_de_la_fonction » donne accès à une aide en ligne détaillée, avec explication sur la fonction, sa syntaxe, des exemples, etc.

6 Mendès France P. (1984). *Pierre Mendès-France. Œuvres Complètes*, 6 tomes, Gallimard. La version numérisée de ces ouvrages, suivant les normes de la *TEI* (*Text Encoding Initiative*) et les recommandations *XML* (*eXtensible Markup Language*), nous a été fournie par le laboratoire *ATILF* (*Analyse et Traitement Automatique de la Langue Française*), *UMR* 7118 *CNRS-Université de Nancy 2*. Les sélections de textes ont été réalisées à l'aide de feuilles de style *XSLT* (*eXtensible Stylesheet Language Transformations*) et de scripts *Perl*.

Ces textes ont été étiquetés morphosyntaxiquement à l'aide du logiciel libre *TreeTagger*[7]. Ils ont été retenus selon une double volonté : (i) permettre une étude des effets quantitatifs sur les cooccurrences en énergie et en disponibilité ; (ii) permettre une comparaison de genres textuels du point de vue de ces cooccurrences. Le corpus s'organise donc autour de 2 genres oraux et 2 genres écrits, avec des textes de taille variable :

- Télégramme (T) : 4 textes ; il s'agit de textes écrits institutionnels adressés par Mendès France dans le cadre de sa fonction ministérielle.
- Note (N) : 4 textes ; ces notes manuscrites sont une sélection de documents de travail internes au ministère.
- Intervention radiophonique (IR) : 4 textes ; il s'agit là de quelques unes des nombreuses interventions radiophoniques (écrit-lu, publique) de P. Mendès France.
- Intervention Parlementaire (IP) : 7 textes ; sélection parmi les nombreuses interventions parlementaires (oral et écrit-lu) du Président du Conseil.

Ces données totalisent 46 808 occurrences, dont la répartition est présentée dans le *Tableau 1*. Nous avons souhaité sélectionner, autant que possible, des textes dont le nombre de mots détermine, selon le même facteur de proportionnalité, le nombre de phrases et de paragraphes. Cette contrainte a deux buts : s'assurer d'une structuration de surface commune pour ces textes ; limiter l'effet d'une variable supplémentaire pour nos cooccurrences, la différence de taille des fenêtres cotextuelles. Dans notre corpus on a donc plus ou moins : Nombre de mots égal à ± 20 fois nombre de phrases, lui même égal à ± 2 fois nombre de paragraphes. Donc, en moyenne, une phrase contient 20 mots et un paragraphe 2 phrases.

3. Taille des textes et profils des nuages énergie-disponibilité

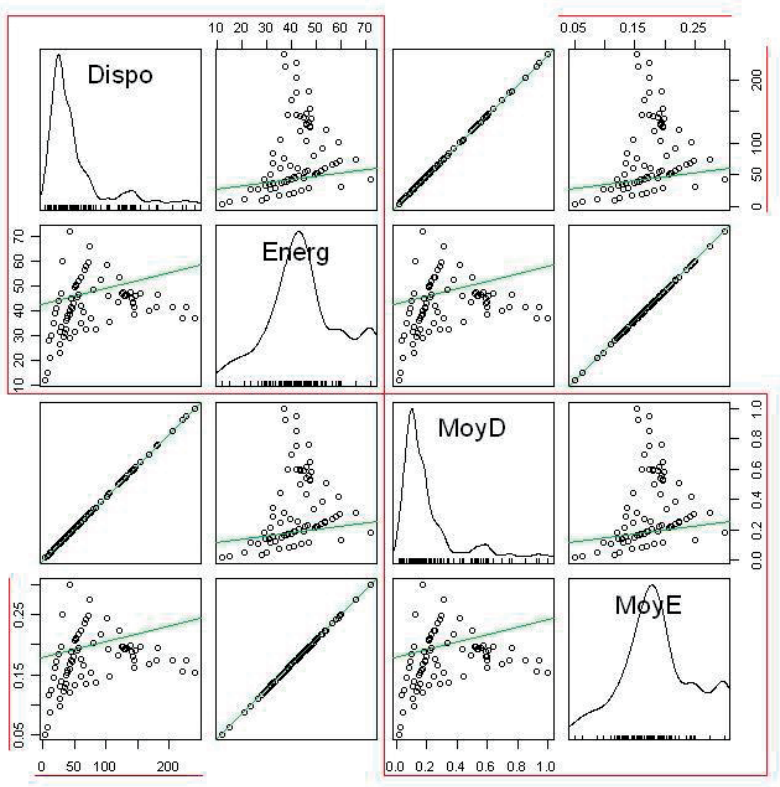
Notre première analyse souhaite mesurer les effets du nombre d'occurrences des textes sur le couple énergie et disponibilité. Pour circonscrire ces phénomènes, nous limitons nos analyses à un genre fixé, le plus nombreux et présentant la plus grande variabilité d'effectifs d'occurrences : l'intervention parlementaire.

Les premiers graphiques (Graphiques 1), nous permettent d'illustrer les différences et points communs entre les couples énergie-disponibilité et énergie-disponibilité moyennes. Nous sommes devant un tableau croisé des graphiques de disponibilité (D), énergie (E), disponibilité moyenne (MD) et énergie moyenne (ME) des lemmes pour le texte IP14 (653 occurrences). Les données des cases en diagonale du tableau intitulées Dispo, Energ, MoyD, MoyE correspondent aux densités respectives de D, E, MD et ME, c'est-à-dire au nombre de mots (en ordonnées) en fonction (en abscisses), respectivement, de D, E, MD et ME. Ces mêmes cases, et leur intitulé, servent à la définition, par leurs directions horizontale et verticale, des axes des graphiques en nuage de points. Ainsi, la première ligne et la deuxième colonne, croisent, dans l'ordre, « dispo » et « énerg », il s'agira donc du graphique des disponibilités du texte (ordonnées) en fonction de son énergie (abscisses). De même, seconde ligne et première colonne croisent, « énerg » et « dispo », il s'agira du graphique des énergies (ordonnées) en fonction des disponibilités

7 *TreeTagger*, *projet TC* (*Textcorpara and Erschliessungswerkzeuge*), *Institut de linguistique computationnelle de l'Université de Stuttgart* : http://www.ims.uni-stuttgart.de/projekte/tc/ et www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

(abscisses). Pour la seconde ligne et la troisième colonne, « énerg » et « MoyD », le graphique a les énergies en ordonnées et MD en abscisse. Etc. On a donc construit des espaces croisant les indices D, E, MD et ME. Dans ces graphiques, chaque point correspond à la représentation d'un lemme du texte en fonction de sa valeur dans les deux dimensions décrites par le graphique.

Parmi ces graphiques, on ne s'intéressera qu'à ceux encadrés en mauve, croisant des indices comparables, à savoir, les énergies et disponibilités des lemmes d'une part, et les énergies moyennes et disponibilités moyennes d'autre part, auxquels s'ajoutent leurs densités respectives. Dans chaque graphique, la droite de régression est représentée en vert. On remarquera, dans un premier temps, que les nuages (E,D) et (ME,MD), et leur densité, sont similaires en tout point, au changement d'échelle près. Pour améliorer la comparabilité des textes étudiés, on préfèrera donc, à partir d'ici, travailler sur le nuage (ME,MD), ramenant les indices E et D, dans un intervalle compris entre 0 et 1.



Graphiques 1 : Nuages de points E-D, ME-MD et densités associées dans le texte IP14

Observons maintenant les densités. Pour ce texte (IP14), le comportement en énergie et disponibilité est très différent. En effet, si l'une et l'autre se profilent comme de relatives gaussiennes, (a) pour l'énergie le maximum de densité (mode), la moyenne et la médiane s'avèrent très voisins (la courbe est centrée), avec une progression en énergie plutôt régulière et bien répartie dans ses valeurs[8], (b) alors que le profil de densité pour la disponibilité est beaucoup plus abrupt, avec un fort pic dans les faibles valeurs de disponibilité et un faible nombre de

8 C'est un profil proche d'une loi normale réduite.

mots pour deux bons tiers des valeurs énergétiques constatées. Il est à noter que l'intervalle des valeurs énergétiques est bien moindre que celui des valeurs de disponibilité, avec un score maximal de 0,3 pour la ME, contre 0,95 pour la MD. (c) L'énergie a donc un profil beaucoup plus « ramassé » que (d) la disponibilité, où la combinaison d'une forte majorité de mots dans les scores faibles et de mots de score très élevé, indique la mise en lumière de lemmes très spécifiques, pour ce texte, du point de vue de leur disponibilité.

La structure du nuage de points rend compte de ces propriétés. (i) On est donc face à un nuage en triangle, avec deux dimensions fortes. L'une est axée autour de l'énergie, l'autre autour de la disponibilité, (ii) constituant le pic du nuage de points pour la disponibilité à proximité du mode des valeurs d'énergie. Ainsi, l'une et l'autre des régressions linéaires (disponibilité-énergie – ligne 1, colonne 2 – et énergie-disponibilité – ligne 2, colonne 1), pratiquement perpendiculaires entre elles, ne parviennent pas à décrire convenablement la forme du nuage, mais la prise en compte de la combinaison des deux, produit une croix qui décrit assez bien cette structure. On remarque, de plus, deux conglomérats de points, (iii) l'un très peu sélectif (c'est-à-dire regroupant beaucoup de mots), correspondant aux valeurs de faible disponibilité autour du mode de l'énergie, qui constitue le centre de la droite de régression disponibilité- énergie, (iv) l'autre, sélectif, situé autour du mode de l'énergie et des valeurs proches de 0,5 et 0,6 pour la disponibilité.

Les Graphiques 2 présentent les nuages de points énergie moyenne-disponibilité moyenne pour les 6 autres interventions parlementaires du corpus. Ils sont accompagnés de leur droite de régression énergie-disponibilité, qui aide à décrire la disponibilité. Les textes-nuages sont présentés par nombre de mots croissants, la fourchette allant de 319 à 16217 occurrences (le texte IP14, avec 653 occurrence s'insérant dans ce classement entre les textes IP16 et IP7, avec, respectivement, 319 et 1464 occurrences).

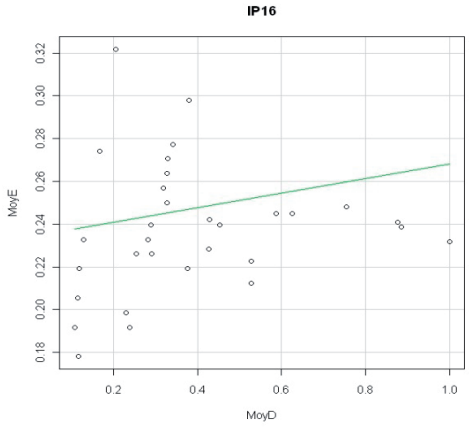
On constate, à l'observation de ces graphiques, que la structure des nuages garde les mêmes deux composantes en énergie et disponibilité que pour le texte IP14, mais de manière de plus en plus stéréotypée à mesure que la taille des textes augmente. On retrouve donc les propriétés décrites précédemment :

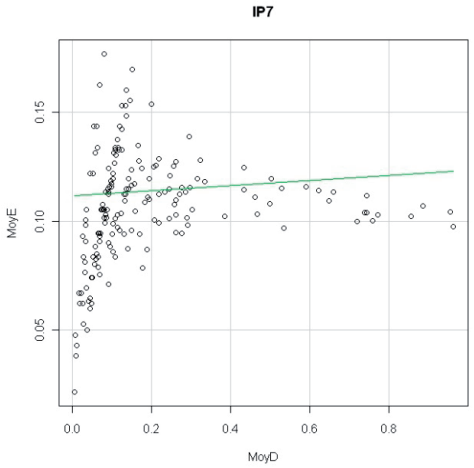
(i) la forme des nuages reste un triangle, qui s'allonge de plus en plus à mesure que la taille des textes augmente, pour s'approcher de deux droites distinctes dès que le seuil des 5000 occurrences est franchi. Ceci est dû au fait que le phénomène (b) se renforce avec la taille des textes, concentrant de plus en plus de mots dans les faibles disponibilités par rapport à l'amplitude des valeurs maximales de celle-ci (qui continuent d'approcher 1 et sont donc de plus en plus rares relativement au nombre total de mots du texte).

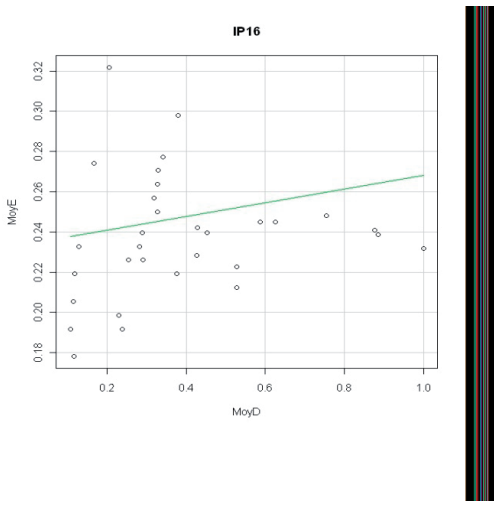
L'énergie, quant à elle, continue sa répartition «normale» (a). Ainsi, on constate encore un pic spécifique de disponibilité pour certains lemmes autour du mode (ou la moyenne ou la médiane) de l'énergie (ii et d).

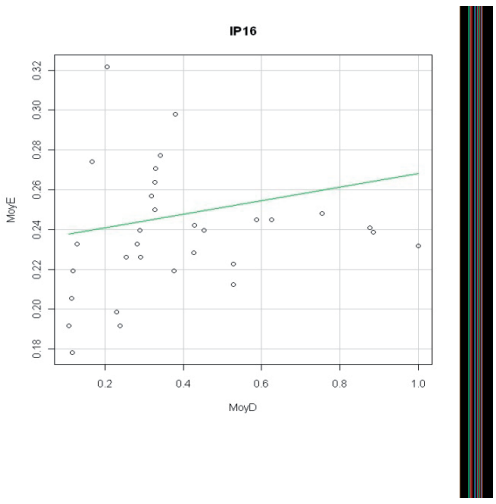
Avec l'augmentation de la taille du texte, l'énergie accentue son profil ramassé (c), avec un maximum de densité de plus en plus faible. On passe d'une énergie moyenne maximale d'approximativement 0,3 pour 319 occurrences, à 0,2 (1464 occurrences) et 0,1 (5066 occurrences), puis 0,05 (plus de 10000 occurrences).

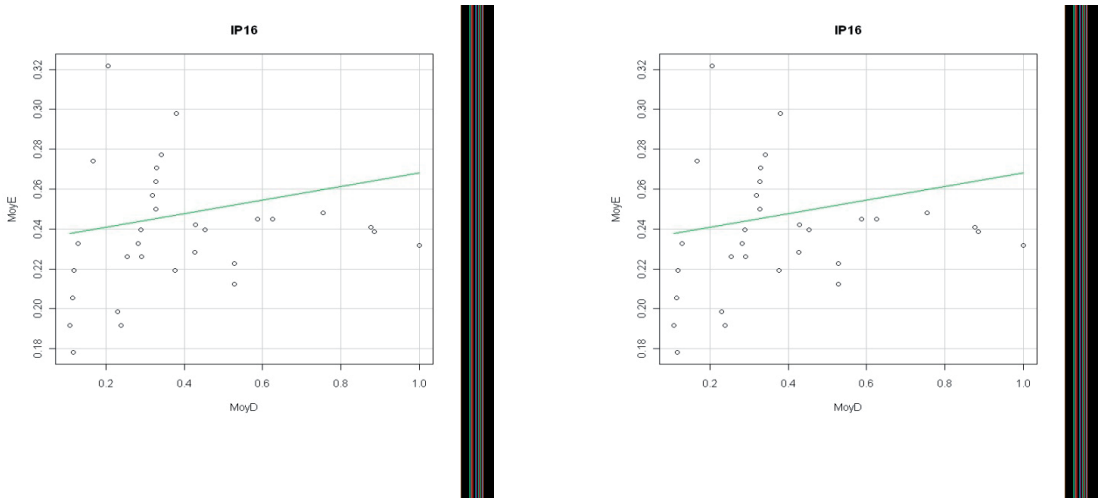
Enfin, si le phénomène d'agglomération des lemmes autour du mode reste constant pour l'énergie (iii), celui décrit pour la disponibilité (iv) ne semble pas apparaître en deçà des 500 occurrences, avec des valeurs variables d'apparition comprises pour la disponibilité, entre 0,3 et 0,8. Si ce phénomène existe bien, il est en tout cas moins marqué que les autres propriétés précédemment décrites. On peut néanmoins se demander, qui sont ces lemmes remarquables et remarqués ? L'analyse qui suit permettra d'éclairer cette question.







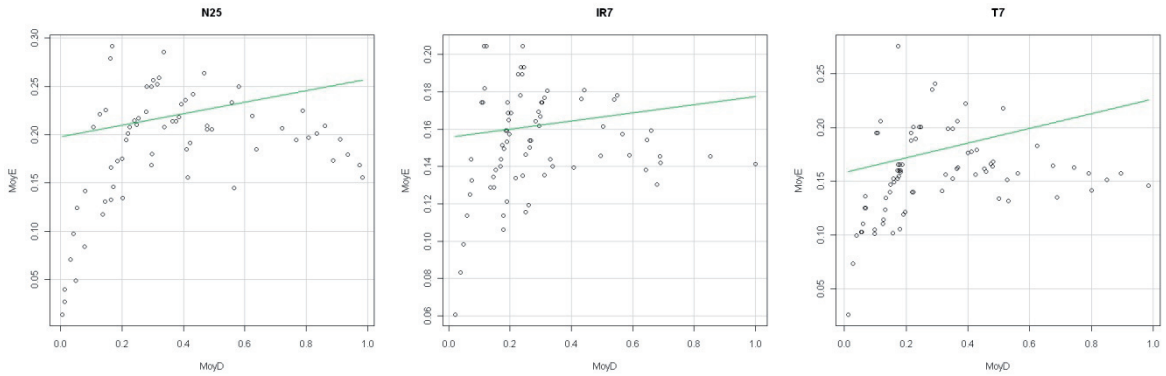




Graphiques 2 : Nuages ME-MD des autres interventions parlementaires, triés par nombre de mots croissant dans les textes

4. Morphosyntaxe et différences en genre

On va maintenant élargir nos recherches en fixant la taille des textes étudiés, mais en ouvrant leur genre. On va comparer les textes N25 (534 occ.), IR7 (624), T7 (651) et IP14 (653). Les Graphiques 3 présentent les nuages de points pour ces textes.



Graphiques 3 : Nuages ME-MD pour les textes N25, IR7 et T7

La forme générale des nuages obtenus diffère peu de celle décrite dans la partie précédente pour le texte IP14 : triangularité, amplitudes, etc. Seule la propriété (iv), déjà mise en doute, n'apparait pas. Pour cause de place, nous ne détaillons pas ici l'ensemble des résultats obtenus sur l'intégralité du corpus, néanmoins, les descriptions faites pour le genre parlementaire restent vraies, en conséquence de la taille des textes, pour tous les autres genres textuels. Doit-on en conclure que l'énergie et la disponibilité ne sont pas des outils satisfaisants pour discriminer de genres textuels ? Avant d'arrêter notre conclusion, nous souhaitons projeter sur ces graphiques la catégorie grammaticale associée à chaque point et réaliser des régressions linéaires énergie-

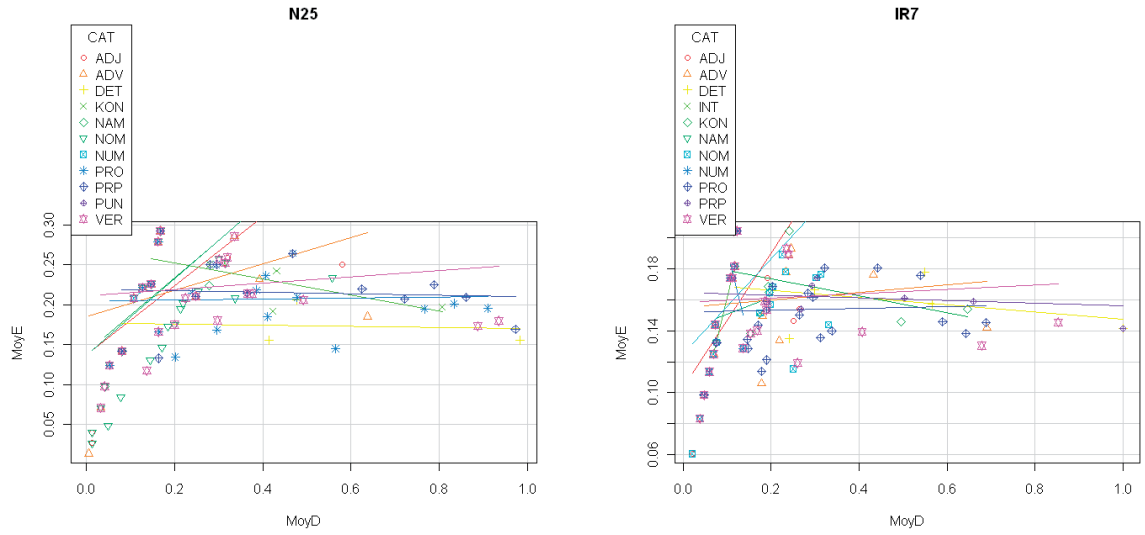
disponibilité indépendantes par classes grammaticales présentes, pour mettre en lumière les différentes parties du discours. Nous avons donc produit les Graphiques 4.

Ces graphiques[9] fournissent de nombreuses informations, que nous n'épuiserons pas ici. On constate tout d'abord quelques constantes dans tous les textes : les mots d'énergie moyenne et forte disponibilité (propriétés ii et d), appartiennent principalement à la classe des «mots outils», qui s'organise en droite à peu près parallèle tout au long de l'axe des abscisses ; les noms, adjectifs, verbes et adverbes sont systématiquement plus proches de l'axe des ordonnées. Ils ont donc une composante disponibilité proportionnellement plus faible que les «mots outils».

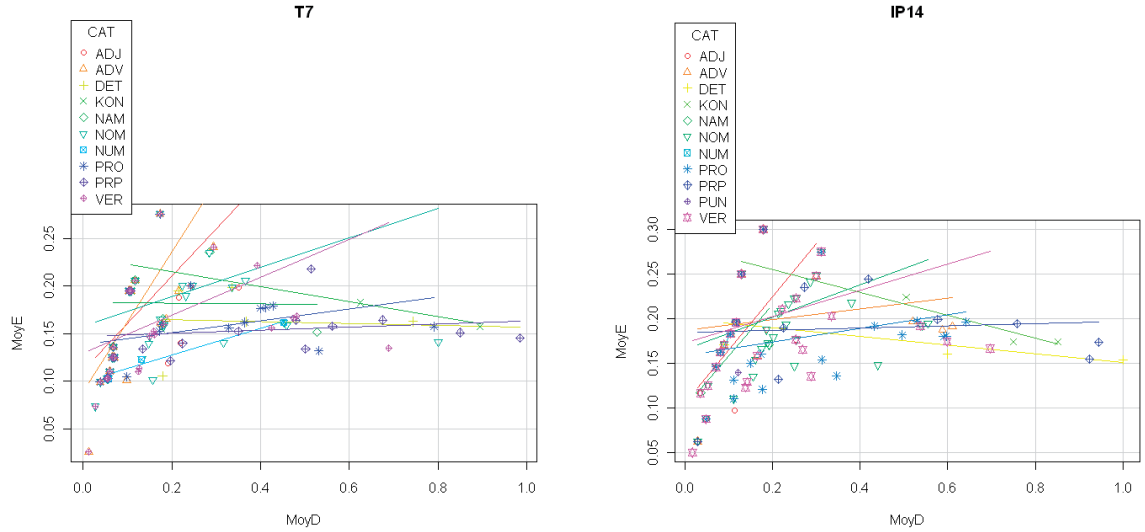
C'est dans l'organisation de ces dernières catégories que l'on observe les principales oppositions d'un genre à l'autre, autour des variations d'attirance pour l'axe des ordonnées (que l'on repère grâce aux droites de régressions). Ces variations semblent moins fonctions des valeurs énergétiques des lemmes associés, que des transgressions de comportement ponctuelles de certains lemmes, dont la position se rapproche de celle des «mots outils», avec une forte disponibilité et une énergie moyenne :

Dans le texte N25, on observe deux verbes de très forte disponibilité (et moyenne énergie), qui rapprochent fortement la droite de régression des verbes d'une parallèle à l'abscisse, tout comme, dans une moindre mesure, quelques adverbes. C'est à nouveau le cas (de façon moins marquée) dans le texte IR7. Le phénomène se répète de manière encore affaiblie pour le texte IP14. Ainsi, sur l'ensemble de ces textes, la partie verbale du discours semble plus «disponible» que la partie nominale.

Mais, dans le texte T7, c'est cette fois un nom qui est touché par l'aspiration d'une forte disponibilité (la droite de régression des noms se rapproche de l'abscisse), même si, dans une moindre mesure, un verbe est lui aussi attiré. Notons que les adverbes et adjectifs ne subissent pas cette attirance.



9 Du fait de la proximité des couleurs, les graphiques sont difficilement lisibles. Nous avons différents moyens de contrôle, grâce au clic sur un point (*cf. supra* note 5) et aux équations des droites de régression.



Graphiques 4 : Nuages ME-MD pour les textes N25, IR7, T7 et IP14, avec mise en valeur des catégories grammaticales

On ne sera pas surpris d'apprendre que dans les textes N25, IR7 et IP14, les verbes à très forte disponibilité sont les auxiliaires *être* et *avoir* et *ne*, *pas* et *plus* pour les adverbes. Quant au nom mis en valeur dans l'intervention parlementaire, il s'agit de *gouvernement*. La disponibilité s'avère être une excellente mesure de la banalité ! Néanmoins, lorsqu'on observe les «mots pleins» de disponibilité moyenne mais s'écartant de la «droite» constituée par ceux de faible disponibilité (iv), on repère des lemmes plus marquants. Dans le texte N25, de nouveau *gouvernement* (MD=0,55) et le verbe *mettre* (0,5) ; dans le texte IR7, *faire* (0,4) ; dans le texte IP14 le verbe *pouvoir* (0,55), *budget* (0,45), *constitution* (0,4) et *décision* (0,3) ; dans le texte T7, les verbes *être* (0,7), *devoir* (0,5), *agir* (0,5), *faire* (0,4), les noms *plan* (0,45), *problème* (0,4), *interlocuteur* (0,35), *esprit* (0,35) et le nom propre *Saigon* (0,55). On obtient ici des marqueurs modaux et thématiques notables (ce qui était déjà le cas de la négation). La transgression vers la disponibilité s'avère être une information précieuse dès lors qu'elle ne touche plus aux termes attendus comme banals. On peut s'interroger sur la similitude de comportement de ces lemmes avec les termes structurant en langue, que sont les «mots outils». Quelle est la part de leur propriété structurante par inhérence avec leur cotexte et celle de leur définition par afférence avec celui-ci ? Dans l'hypothèse ouverte par X. Luong (Luong *et al.*, 2010 : 330-331) d'une énergie marqueur du comportement en langue et d'une disponibilité du comportement en discours, cette double transgression (du nuage et de nos attentes) pourrait être la marque d'un fonctionnement en langue de structuration du discours.

Nous avons prolongé nos analyses à l'ensemble du corpus. La mise en avant régulière par une forte disponibilité des auxiliaires *être* et *avoir* se retrouve dans des textes de tous genres textuels. On note aussi des décrochages de noms ou de verbes en position de disponibilité moyenne (entre 0,3 et 0,8). A titre d'exemple, on peut noter que les interventions parlementaires et les télégrammes se distinguent des autres textes par l'apparition dans cette position, en décrochage par rapport aux autres verbes, des verbes modaux (*devoir*, *pourvoir*, *savoir*) et par l'attraction

de noms dans cette même zone, ce qui est peut-être une marque du discours politique public, par opposition aux productions privées.

5. Conclusion

Les nuages de points, avec l'accroissement de la taille des textes, se caricaturent. Ce qui devient un biais lorsque l'on veut comparer des données de tailles différentes. Particulièrement en envisageant un traitement automatique des textes. Néanmoins, notre étude, même si elle ne s'intéresse qu'à 4 genres et aux discours d'un seul individu, nous a permis de mettre en valeur un certain nombre de constantes et leurs implications.

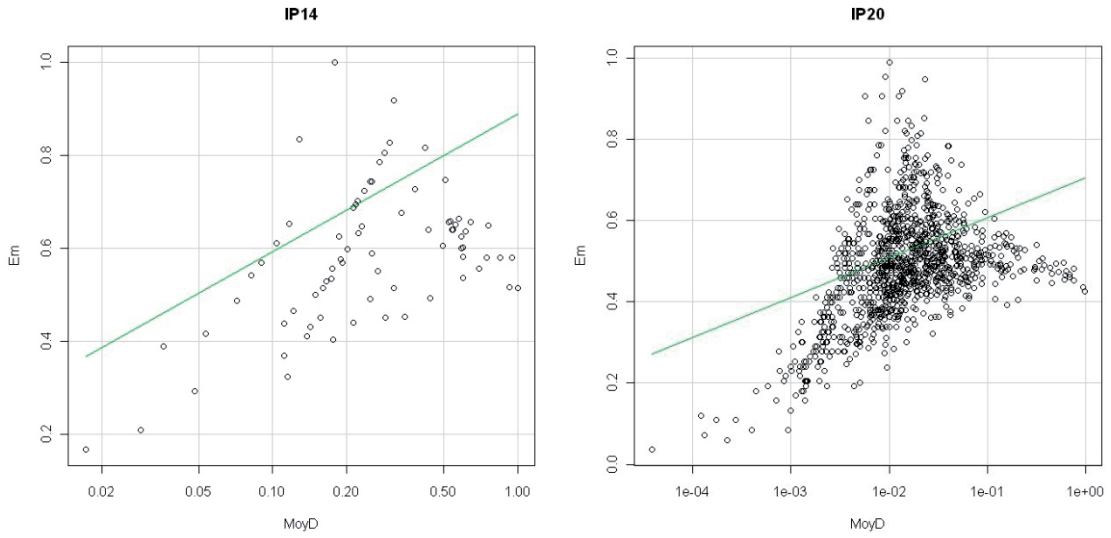
On a remarqué que, sensible au nombre d'occurrences de la forme et à sa répartition dans le texte, la disponibilité est une mesure très efficace de la banalité. On a ainsi pu constater que les «mots outils», et *être* et *avoir*, présentaient des positions extrêmes et stables, que nous apparentons à du «bruit» pour nos analyses. Nous pensons donc que ces parties du discours, y compris les auxiliaires *être* et *avoir*, doivent être écartées. De même, les auxiliaires modaux se sont révélés comme une classe particulière qui justifie un étiquetage spécifique. Il est à prévoir que le jeu d'étiquettes le plus efficace dépend du corpus. En l'état, on ne peut donc pas éviter, en amont de tout traitement automatique, une étude «à la main» dans le but de construire les meilleures étiquettes des parties du discours pour notre corpus.

En faisant l'hypothèse que les textes, d'un même auteur et d'un même genre textuel, possèdent la même structure interne, on attendait, dans le point 3 de notre exposé, que les nuages énergie-disponibilité produits soient de même forme. Vue la régularité des transformations produites par la taille sur la forme des nuages, nous souhaitons soumettre un certain nombre d'adaptations pour les contourner et ouvrir des pistes pour une comparaison automatique des textes entre-eux.

Comme on l'a vu, l'intensité de l'énergie est fonction de la taille du texte (c). Ainsi, pour comparer deux textes de tailles différentes, sans subir ce resserrement, un moyen est de ramener l'énergie à l'unité, en utilisant la valeur maximale de l'énergie moyenne des mots du texte. On obtient, pour un texte A, de lemmes $(m_i), \overline{E_A^I}(m_j)$ l'énergie moyenne du mot m_j :

$$\overline{E_A^I}(m_j) = \frac{\overline{E_A}(m_j)}{\max(m_i)} = \frac{E_A(m_j)}{\max(m_i) \times \sqrt{N_A - 1}}$$

De plus, l'utilisation du logarithme sur la disponibilité peut permettre de limiter la déformation (i). Testons ces aménagements sur les textes IP14 (653 occ.) et IP20 (12047) (Graphiques 5, notation Em pour l'énergie moyenne) : l'amplitude des énergies est bien comprise entre 0 et 1. De plus, bien que de densités très différentes, les deux nuages ont des formes assez proches, qui rappellent leurs formes antérieures avec un recentrage des valeurs énergétiques vers le centre du graphique. Reste à opérer un changement de variable sur l'axe des abscisses pour normer ce nuage entre 0 et 1 en tenant compte des déformations de l'échelle logarithmique. On voit ainsi apparaître des valeurs d'énergie et disponibilité moyenne, mais aussi des valeurs de forte énergie et de disponibilité moyenne.



Graphiques 5 : Nuages de points modifiés pour les textes IP14 et IP20

Appliquée aux catégories grammaticales, notre méthode s'avère d'ores et déjà efficace pour repérer des lemmes spécifiques des textes. On a, de plus, discuté de façon avancée la possibilité de comparer des textes de taille variable. Il reste néanmoins un grand pas à franchir pour envisager une comparaison automatique de textes : déterminer l'indice pertinent pour décrire la forme des nuages ; en complexifiant encore la question si l'on souhaite tenir compte des diverses catégories morphosyntaxiques. Ce qui semble une nécessité.

Références

Brunet E. (2003). Peut-on mesurer la distance entre deux textes ? In *Corpus*, n°2. http://corpus.revues.org/index30.html.

Charolles M. (1997). L'encadrement du discours, univers, champs, domaine, espaces. In *Cahier de recherche linguistique*, n°6, Université Nancy 2, pp. 1-73.

Fox J. (2004). The R Commander: A Basic-Statistics Graphical User Interface to R. In *Journal of Statistical Software*, Vol. 14, n°9, pp.1-42.

Luong X. *et al.* (2010). La cooccurrence, une relation asymétrique ? In *Actes des JADT 2010*, Vol.1, Rome, pp. 321-331.

Martinez W. (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*. Thèse de doctorat, Université Sorbonne Nouvelle – Paris III.

Viprey J.-M. (1997). *Dynamique du vocabulaire des Fleurs du mal*, Honoré Champion.