**Classification de champs sémantiques en partitionnement de données : une approche expérimentale**

Joé Bourgeois-Paquin1, Jean-Simon Bondaz1 et Claudia Roy1

1 Département d’informatique, Cégep du Vieux Montréal, Montréal, Québec

Dans le cadre du cours de Projet Oracle – B62 (Hiver 2019)

Résumé

Cet article présente des pistes d’analyse de données organisées en partionnement (*data clustering*), celles-ci étant recueillies principalement dans des corpus de texte en français. Ces sous-ensembles sont peuplés par des individus (mots) dont la coocurrences dans les corpus avec les autres mots est comptabilisée. Le vecteur de coocurrences de chaque individu sert ensuite de coordonnées permettant de calculer sa distance à un barycentre choisi ou généré aléatoirement Cette distance nous permet de former des sous-groupes représentant des champs sémantique basés sur leurs cooccurences entre invididus.

Mots-clés: *partionnement, cooccurence, corpus*

1. Introduction

Le présent article est le fruit du cours de Projet Oracle (B62) dans lequel nous avons été encouragés de faire l’analyse de corpus de texte en français par une intelligence artificielle programmées à lire les données, les extraire dans une matrice de cooccurences avec les autres données et entrainer ce modèle à former des sous-groupes par champ sémantique.

1.1 Problématique

Cet outil nous a amené à nous demander s’il était possible d’analyser ces sous-groupes d’invidivus pour en induire un sens dans la classification systématique de notre intelligence artificielle, principalement en contrôlant trois potentielles variables : quantité de partitions, thème des corpus et classification par syntagme verbal (dans l’optique morphologique de ce qu’est un verbe et ses déclinaisons).

1.2 Hypothèses

1.2.1 Quantité de partitions (centroïdes)

Notre première hypothèse repose sur l’idée que le nombre de centroïdes fera varier le champ sémantique pour former des sous-groupes plus cohérents à l’échelle sémantique. Nous croyons qu’un trop petit nombre créera de trop grands groupes « pêle-mêles » ayant peu de cohérence et un trop grand nombre de sous-groupes divisera trop amplement les sens au point où les liens entre les individus s’éffriteront. Nous donc sommes d’avis que la cohérence des champs sémantiques suit une parabole tel que :

*X = Nombre de centroïdes*

*Y = Partitions ayant une relative cohérence à l’échelle sémantique.*

1.2.2 Thème des corpus

Notre deuxième hypothèse est l’idée que les sous-groupes deviendront davantage cohérents si les corpus ont en commun un thème principal. Nous croyons qu’il y aura plus de chance de redondance de mots, augmentant ainsi globalement le nombres de cooccurrences entre les individus. Nous croyons que ce critère augmentera la cohérence sémantique des sous-groupes et que cela sera effectif peu importe la taille la fenêtre d’analyse du curseur d’entrainement tel que :

*X = Taille de la fenêtre d’analyse*

*Y = Partitions ayant une relative cohérence à l’échelle sémantique.*

1.2.3 Classification des champs par groupes verbaux

Notre dernière hypothèse repose dans l’idée qu’il est possible que les individus aient tendance à se regrouper par critères propres à un groupe syntaxique. Dans notre cas, nous croyons que les verbes dans les mêmes sous-groupes auront en commun les catégories verbales soit : le mode, le temps et la personne grammaticale tel que :

Y = Niveau de cohérence sémantique par critères  
X = Catégories verbales (Mode, temps, personne)

Nous croyons que cette hypothèse est particulièrement intéressante parce qu’elle tend à décortiquer un critère propre à un échantillon de la population ayant en commun une catégorie grammaticale (les verbes) et le possible lien avec sa cooccurence avec les autres verbes. Par exemple, si une phrase utilise principalement l’indicatif présent, les autres verbes dans la même fenêtre ont de fortes chances d’avoir ce mode et d’avoir ce temps.

2. Matériel et méthodes

2.1 Matériel

Les corpus sur lequel nous nous attèleront, en premier lieu, sera 40 000 mots uniques (certains même en anglais) issus de la littérature française (Albert Camus, Alexandre Dumas, etc.) et des textes du psychothérapeutre Alfred Adler.

Ce matériel sera lu avec un curseur en utilisant une taille de fenêtre variable pour calculer la cooccurrence des mots avec leurs voisins syntaxiques. Cette fenêtre, pour l’ensemble de ces textes, restera de taille sept, puis 5, 7 et 10 pour l’analyse des textes d’Adler de manière isolée. Ces mots (ou individus) seront ensuite comptabilisés en base de donnée ainsi que leur cooccurence avec les autres mots selon la taille de fenêtre.

Pour l’analyse verbale des corpus, nous utiliserons une base de donnée de lexique en français pour déduire leur catégorie grammaticale, triant ainsi seulement les verbes et ainsi voir leur mode, leur temps et la personne à laquelle il a été conjugué (ou non dans le cas d’un lemme comme un verbe à l’infinitif).

2.1 Méthodes

Les méthodes pour classifier ces individus se baseront sur l’approche de partitionnement de données laquelle déduira des catégories sémantiques semblablent à des champs lexicaux. Celle-ci nous permettra de classifier ces mots par calculs statistiques afin d’y tracer de potentiels champs lexicaux.

Tout d’abord, nous générerons des centroïdes à partir de mots entrés, desquels nous attribuerons des coordonnés par rapport aux cooccurences à d’autres mots. Cet ensemble de données nous permettra, d’une part, de générer des centroïdes en les positionnant sur certains de ces individus ou, d’autre part, de générer des centroïdes ( aléatoirement tel que sa coordonnée () suit ces règles :

Nous générerons un unique vecteur () dans lequel chaque index représentera un individu et la valeur à cet index sera le centroïde () auquel il appartient, c’est-à-dire, celui qui est le plus proche (selon la formule du moindre carré) :

Une fois tous les individus associés à un centroïde, nous repositionneront celui-ci à son barycentre dans la zone attribuée en effectuant la moyenne de positionnement du centroïde selon :

Une fois les centroïdes repositionnés, nous réaffecterons les individus aux nouveaux centroïdes et compileront ceux qui ont effectués un changement de zone; l’absence de changement de zone pour tous les individus est la condition d’arrêt des itérations de repositionnement des centroïdes. Une fois condition d’arrêt de l’itération principale de l’algorithme comblée, nous assumerons que les partitionnements sont complets.

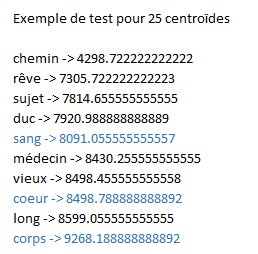
Pour tester notre troisième hypothèse[[1]](#footnote-2), les partitions formées seront triées à partir du lexique de mots en français en ne conservant que ceux qui sont potentiellement des verbes.

3. Résultats

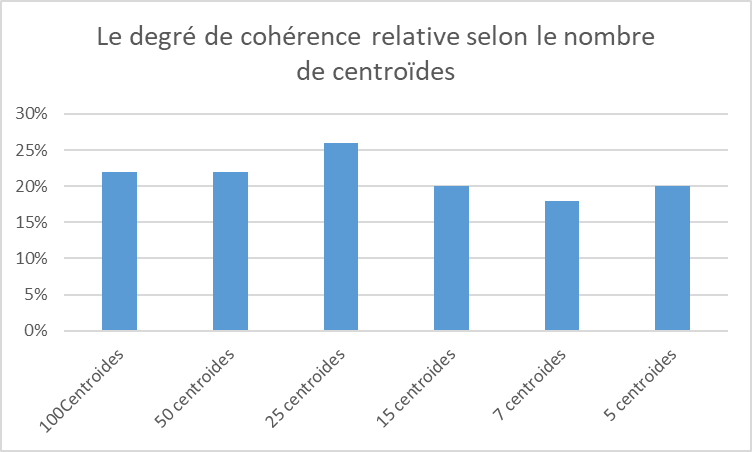
3.1 Quantité par partitions

Pour chaque test, nous avons choisi cinq centroïdes au hasard et nous avons analysé la cohérence relative de chaque centroïde avec lesquels nous en avons fait une moyenne. Ainsi, nous nous assurions d’avoir une meilleure vue d’ensemble de tous nos tests. De plus, nous nous sommes assurés que les centroïdes que nous avions choisis avaient tous dix mots d’affichés. Par contre, lorsqu’il était question de faire des tests avec moins de centroïdes, nous avons dû prendre les partitions qui n’avaient pas nécessairement dix mots. Chacun des tests a été effectué avec la même banque de mots et avec la même taille de fenêtre.

*Exemple de sélection du degré de cohérence relative*



Histogramme du degré de cohérence relative selon le nombre de centroïdes



3.2 Thème des corpus et taille de fenêtre

Similairement au test précédent (Quantité par partions), nous avons choisi cinq centroïdes aléatoirement pour chaque donnée et nous avons analysé la cohérence relative du contenu de chacun d’eux. La donnée finale est le résultat de la moyenne des cinq centroïdes analysés. Nous avons choisi d’analyser l’angle de la relation entre la cohérence relative des centroïdes et la taille de la fenêtre avec laquelle les cooccurrences ont été mesurées. Nous avons aussi trouvé pertinent d’inclure en supplément le nombre de centroïdes avec lequel le test a été fait. Ils ont d'ailleurs été faits à partir de textes traitant du même sujet, soit la psychologie humaine.

Histogramme du degré de cohérence relative selon la taille de la fenêtre et nombre de centroïdes

3.3 Classification des verbes

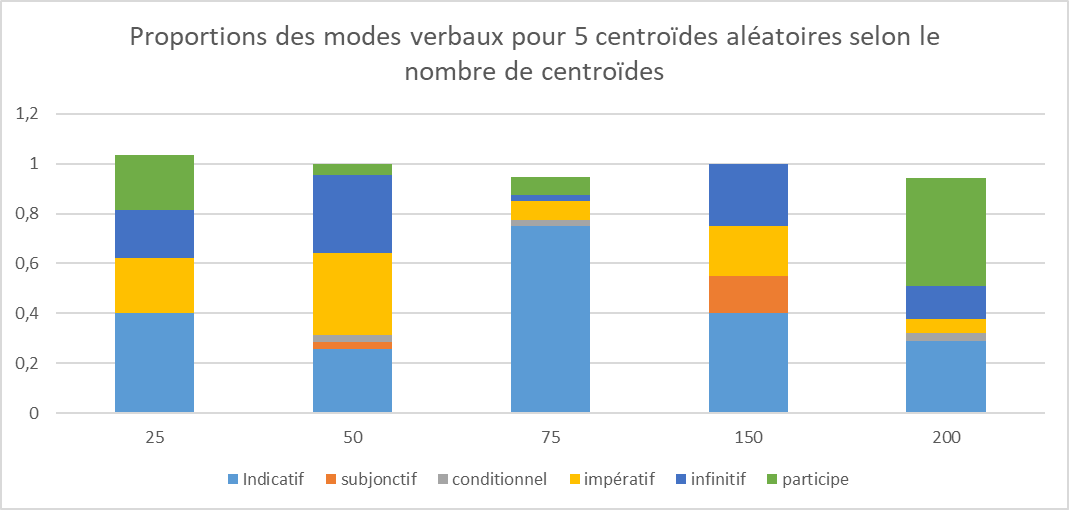
Les verbes ont été analysés sous trois angles : le mode, le temps et la personne grammaticale. Les résultats ont été divisés par la quantité de partitions employées (25, 50, 75, 150 et 200) et mesurés selon leur proportion dans l’ensemble de l’échantillon analysé, soit une taille d’échantillon de 5 centroïdes dans tous les cas, et en mesurant le moyenne parmi le nombre de mots affichés par centroïdes.

Par exemple, s’il y avait trois verbes ayant le mode « Indicatif » sur les quinze verbes les plus près du centroïde, la proportion du mode indicatif serait donc de 0,2 (20%) dans cette partition :

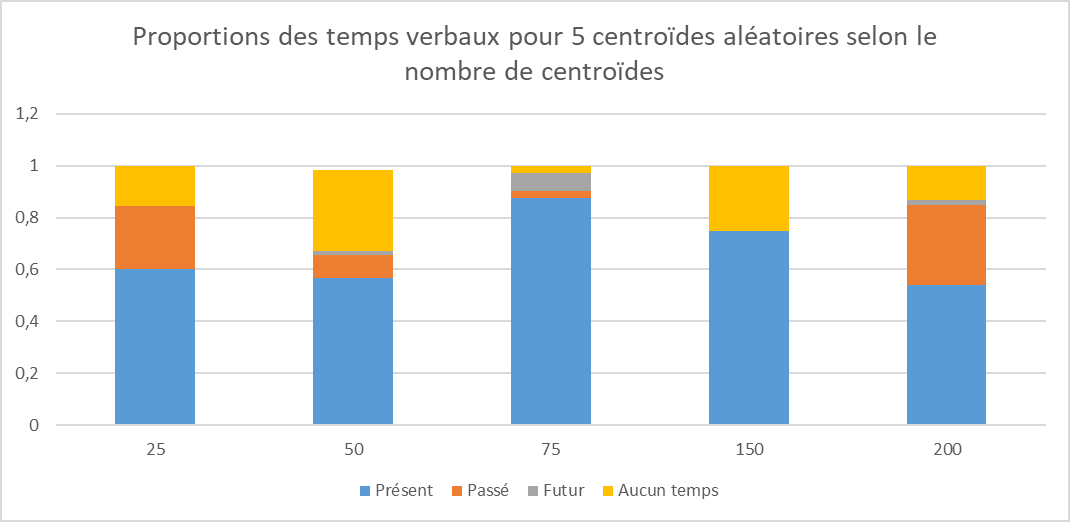
[[2]](#footnote-3)

Dans tous les histogrammes, la variable dépendante est représentée par la proportion de la variable étudiée sur l’ensemble sur 1. Par exemple, une valeur de 0,5 serait l’équivalent de 50% de proportion pour une telle valeur sur l’ensemble des autres individus.

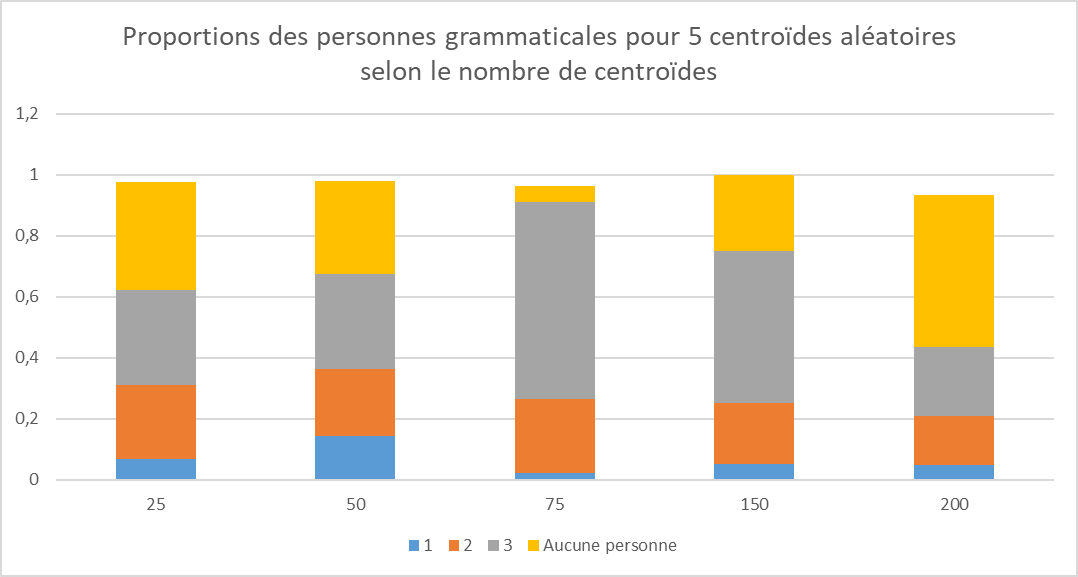
Histogramme des proportions des modes verbaux pour 5 centroïdes aléatoires selon le nombre de centroïdes



Histogramme des proportions des temps verbaux pour 5 centroïdes aléatoires selon le nombre de centroïdes



Histogramme des proportions des personnes grammaticales pour 5 centroïdes aléatoires selon le nombre de centroïdes



4. Discussions

4.1 Quantité par partitions

Selon l’hypothèse que nous avions émise, plus il y a de centroïdes plus nous devions avoir un haut pourcentage de degré de cohérence relative. Avec les résultats que nous avons obtenus, nous observons qu’un trop petit nombre de centroïdes créé de trop grandes partitions sans trop de cohérence sémantique. Le même phénomène est observable quand il y a un grand nombre de centroïdes (20% à cinq centroïdes et 23% pour cent centroïdes). Nous concluons toutefois que pour un affichage de dix mots par partition, le degré de cohérence relative est assez stable, c’est-à-dire que la différence entre un paramètre de cinq centroïdes et de cent centroïdes est relativement négligeable.

4.2 Thème des corpus et taille de fenêtre

Notre hypothèse de départ stipulait que si nous faisions les tests de centroïdes à partir de textes ayant le même sujet, nous pourrions remarquer un degré de cohérence plus élevé que dans des textes n'ayant aucun lien apparent entre eux. Nous remarquons que malgré le fait que les mots partagent les même champs lexicaux, il est difficile d’établir un lien concluant entre les individus du même centroïde. En effet, pour une fenêtre de taille de 5, le degré de cohérence se situe entre 8% et 18%, pour celle de taille de 7 le degré de cohérence est entre 3% et 26% et pour une taille de fenêtre de 10, le degré de cohérence varie entre 12% et 32%. Nous constatons que les résultats sont trop distants pour pouvoir affirmer qu’avec un échantillon de mots provenant du même champ lexical il y ait une cohérence relative significative.

4.3 Classification des verbes

Notre hypothèse de départ pour cette partie de l’expérience était de voir si les individus se regroupaient selon leurs caractéristiques syntaxiques lesquelles étaient le mode, le temps ou la personne grammaticale des verbes.

Tout d’abord, dans les résultats sur le mode des verbes, nous observons une répartition qui semble naturelle des modes de verbes. En effet, nous croyons que la prépondérance en nombre des indicatifs dans un texte s’explique par un choix de cohérence grammaticale par l’auteur, au même titre que les autres modes. Toutefois, nous observons aussi une énorme augmentation du nombre relatif de verbe à l’indicatif lorsqu’il y a soixante-quinze partitions.

Notre théorie face à ce phénomène est qu’un tri plus efficace des noms s’effectue à ce nombre de partitions. Étant donné que nous n’avons aucun moyen de déterminer si l’individu est un verbe à l’indicatif présent à la troisième personne en ou un nom masculin (les deux ayant très souvent la même graphie en français), nous croyons que le phénomène faux notre analyse des verbes.

Sur la base de notre échantillon et de la faiblesse de sa fiabilité, nous sommes contraints de réfuter notre hypothèse de départ et d’affirmer qu’il nous est impossible de cerner un lien entre les critères grammaticaux des verbes et de leur coocccurence avec les autres verbes.

5. Conclusion

Les résultats récoltés s’avèrent peu concluant par rapport aux hypothèses que nous avions posé au départ. Nous croyons qu’il aurait été possible d’améliorer les tests pour obtenir des résultats plus significatifs. Nous aurions pu utiliser un bassin de textes plus grand, permettant ainsi d’avoir beaucoup plus de données. Il nous aurait été aussi possible d’optimiser notre code, en explorant d’autre méthodes de clustering peut-être plus efficace.

Remerciements

Nous remercions notre enseignant Pierre-Paul Monty de nous avoir aidé dans l’élaboration, dans la mise en place et de son appui dans cette expérience.

1. Voir *« 1.2.3 Classification des champs par groupes verbaux »* [↑](#footnote-ref-2)
2. « Proportions des modes verbaux pour 5 centroïdes aléatoires selon le nombre de centroïdes. » [↑](#footnote-ref-3)