Analyse Graphique K-500

Trois K-500 ont été effectués, soit un clustering pour chaque taille de fenêtre 5, 7 et 9. Parmi les résultats, les 100 groupes (clusters) de chacun ont ensuite été sélectionné aléatoirement comme échantillon, dans le but d’analyser l’homogénéité des types de mots regroupés ensemble. Si 60% des mots d’un même groupe appartenaient au même type mots (déterminants, verbes, adjectifs, etc.), l’occurrence du type de mots en question était incrémenté de 1. Lorsqu’un cluster était trop mixte pour déterminer une tendance, la catégorie « Impossible à déterminer » était incrémentée.

Le graphique ci-dessous illustre la moyenne d’occurrence de chaque type de mots à l’intérieur des trois échantillons. Visuellement, les résultats sont assez clairs : les clusters les plus homogènes contenaient plus souvent des verbes (34.67%) et des noms (31%). Il était également impossible à déterminer une tendance précise dans 22% des groupes. La courbe orange du graphique, illustrant le total des moyennes sur 100%, montre que ces trois catégories à elles-seules constituent presque 90% des résultats. Notre hypothèse est que cela est dû à la structure de la langue française elle-même, les phrases étant construites et ordonnées la forte majorité du temps en sujet, verbe et complément. Les verbes et les noms se retrouvent ainsi souvent à la même position dans une phrase. La présence, quantité et position d’adjectifs, adverbes, etc., peuvent, quant à eux, varier selon la phrase ou la prose de l’auteur, ce qui pourrait sans doute expliquer le nombre de groupe mixtes « Impossible à déterminer ».

Analyse groupes de mots

Au-delà de l’homogénéité des groupes de mots selon leur type, nous réalisons que l’algorithme du k-means réussit également à grouper des mots selon une expression ou un champ lexical rapproché. Dans les résultats du K-500 avec taille de fenêtre 5 par exemple, le groupe 222 contient [six, vingt, cinq, huit, dix, ans], possiblement dû à l’énumération du temps (ex : « six ans plus tard ») ou lorsque l’âge des personnages est mentionné dans les textes. Le groupe 332 [toutes, hommes, mains, tous, autres, yeux, trois, deux] est également intéressant. À première vue, il semble un peu hétéroclite, même si quelques pairs de mots se ressemblent lexicalement (toutes/tous, mains/yeux, trois/deux). Si l’on considère toutefois la façon dont les mots sont agencés dans une phrase, ce groupe de mots prend alors tout son sens : « tous les autres », « tous les hommes », « toutes les trois », « tous les deux », « tous les yeux », etc.

L’algorithme regroupait également parfois plusieurs mots évoquant ensemble un champ lexical commun ou un certain lieu. Par exemple, le groupe 312 [caisses, plats, ventres, fromages, énormes, vagues, carottes, travailleurs, roses, bruits] pourrait faire référence à un marché ou la place publique, où des marchands exhibent d’ « énormes caisses » de « fromages », de « carottes » ou de « roses », des restaurants prêts à servir leurs « plats » à des « vagues » de « travailleurs » affamés aux « ventres » vides. Il en est de même pour le groupe 234 [port, jardin, canal, salon, bastion, bureau, mercier, corridor, sol, marché], dont les mots représentent tous des lieux (sauf le mot « sol », qui aurait pû toutefois décrire l’état des lieux). Ce sont tous des mots qui auraient pu décrire une ville/village, ou préciser les déplacements d’un personnage, tel un « corridor » menant au « bureau » ou « salon », un « canal » menant au « port », se rendre chez le « mercier » au « marché », etc.