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 ».