Le Dictionnaire Électronique des Synonymes (DES) et ses graphes d'adjacence

Laurette Chardon

Le Dictionnaire Électronique des Synonymes(DES) et ses graphes d'adjacence

PAR LAURETTE CHARDON · 06/05/2020

Introduction

Développé dans les années 1990 par deux chercheurs et mis en ligne au début des années 2000, le Dictionnaire Électronique des Synonymes ou DES du laboratoire CRISCO est un projet de recherche utilisant des outils mathématiques complexes pour représenter la synonymie. La base de départ, constituée à partir de sept dictionnaires classiques issu de l'INALF (Institut National de la Langue Française) a été améliorée avec un important travail de correction. Une vidéo de présentation de quelques minutes contient les éléments essentiels à retenir et une présentation détaillée du DES est disponible pour ceux qui souhaitent en savoir plus.

Le DES a régulièrement connu une fréquentation croissante des internautes : il est consulté à ce jour de 150.000 à 200.000 fois par jour soit environ deux requêtes par seconde. Bien sûr, il serait inexact d'extrapoler en affirmant qu'il s'agit de 150.000 à 200.000 visiteurs par jour ! Il est évident que très souvent un même visiteur effectue plusieurs requêtes comme nous le constatons dans les deux vidéos de remerciements [Note2] .

Comment expliquer un tel engouement ? Le premier argument en faveur du DES est sa simplicité d'usage : son interface est sobre et sans publicité. Les second et troisième arguments concernent la régularité de sa mise à jour et la facilité de passage d'un mot à un autre en un seul clic grâce à des liens hypertextes. Enfin le quatrième argument et non le moindre est sa « synonymie élargie »[Note1] facilitant le travail de reformulation des rédacteurs.

Il existe une autre partie du DES, moins connue et certainement moins facile à appréhender : il s'agit de l'**espace sémantique**. Plusieurs articles ont déjà été réalisés sur ce sujet : un tutoriel avec *curieux*, les lettres d'actualités du DES n°7 avec ***travail***, n°6 avec ***gagner*** et n°5 avec ***responsable*** et enfin un article sur ce blog « Le Dictionnaire Électronique des Synonymes du CRISCO et l'éventail des sens lexicaux » avec ***comprendre, compter, entendre et importer***. L' intérêt de l'espace sémantique est d'obtenir rapidement sous forme graphique les différents sens d'un mot.

Ce post a pour but d'une part de donner quelques éléments de statistiques sur le DES et d'autre part de détailler un outil graphique autre que l'espace sémantique mais ayant la même finalité : le **graphe d'adjacence.** Avec cet outil, nous aborderons une méthode simple de nettoyage afin de rendre ce graphe plus lisible pour les vedettes qui ont beaucoup de synonymes.

Mais tout d'abord commençons par quelques statistiques caractéristiques du DES.

Quelques statistiques

Le DES possède à ce jour plus de **50.000 entrées et près de 210.000 liaisons synonymiques**. Pour chaque entrée, nous obtenons, entre autres, la liste des synonymes.

⇒ Nombre de synonymes

Une première information statistique nous est donnée par la figure 1, ci-dessous, qui nous montre le nombre d'entrées du DES en fonction du nombre de synonymes.

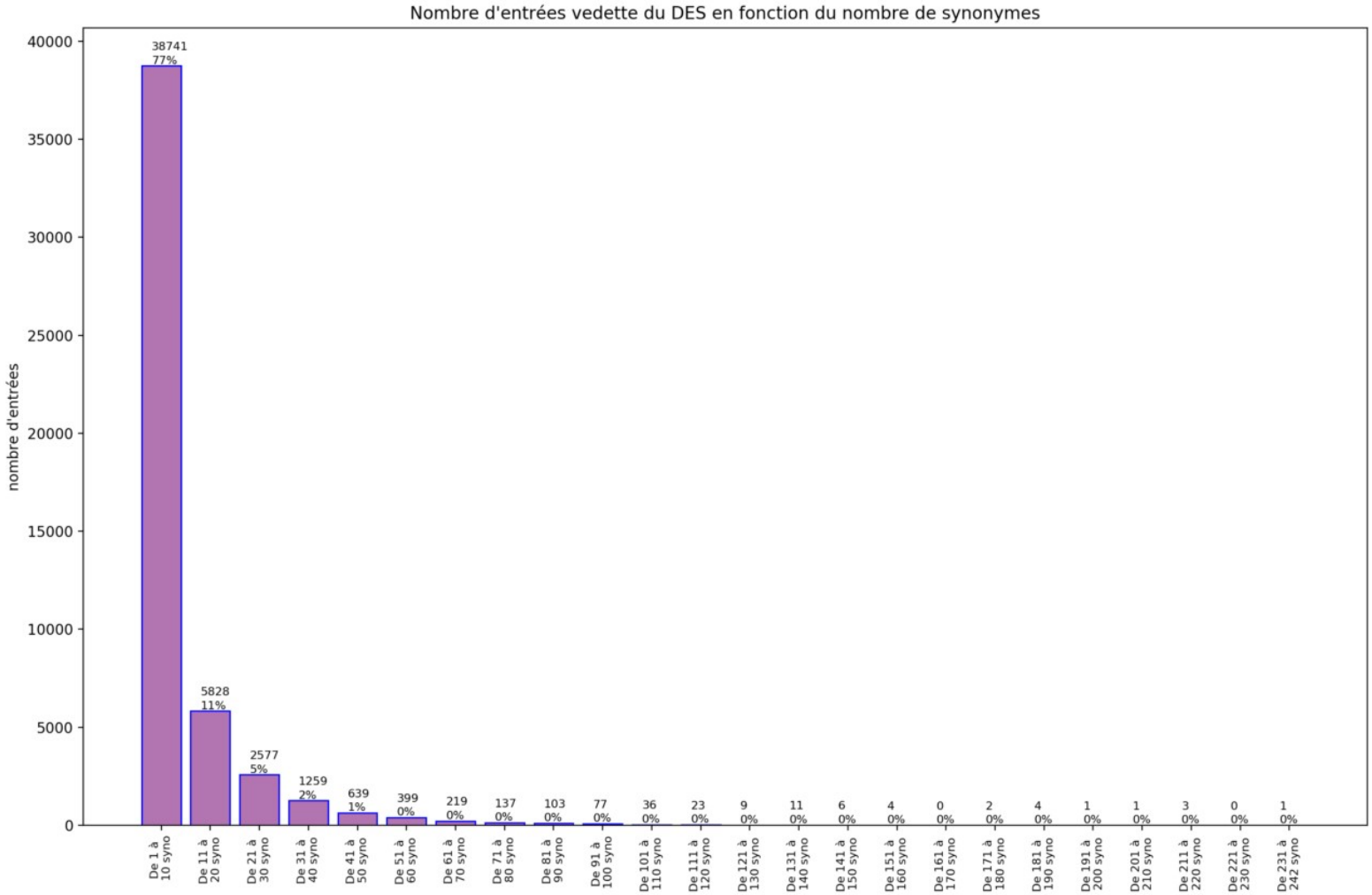


Figure 1 : Nombre d'entrées du DES en fonction du nombre de synonymes (données de mars 2020)

Nous constatons qu'une très grande majorité **77%** soit 38.741 **ont moins de 10 synonymes** et nous tombons très vite à 11% (soit 5828) des entrées qui possèdent de 11 à 20 synonymes. Seules quelques entrées ont un très grand nombre de synonymes. Par exemple, nous avons en partant de la droite du graphique : ***bon*** (242 synonymes),

faire (219), ***fort*** (215), ***prendre*** (212), ***bien*** (205), ***beau*** (193), ***passer*** (190), ***abri*** (184), ***battre*** (183), ***mauvais*** (183), ***diminuer*** (180) et ***fou*** (176) pour ne citer que les plus importants. Certains comme ***fort*** et ***bien*** inclut la forme adjectivale comme nominale.

Si, de plus nous zoomons sur les 77 % des entrées qui ont moins de 10 synonymes, nous voyons finalement, dans la figure 2 ci-dessous, que **64 % n'ont que de 1 à 5 synonymes** et cela tombe à 13 % entre 6 et 10 synonymes.

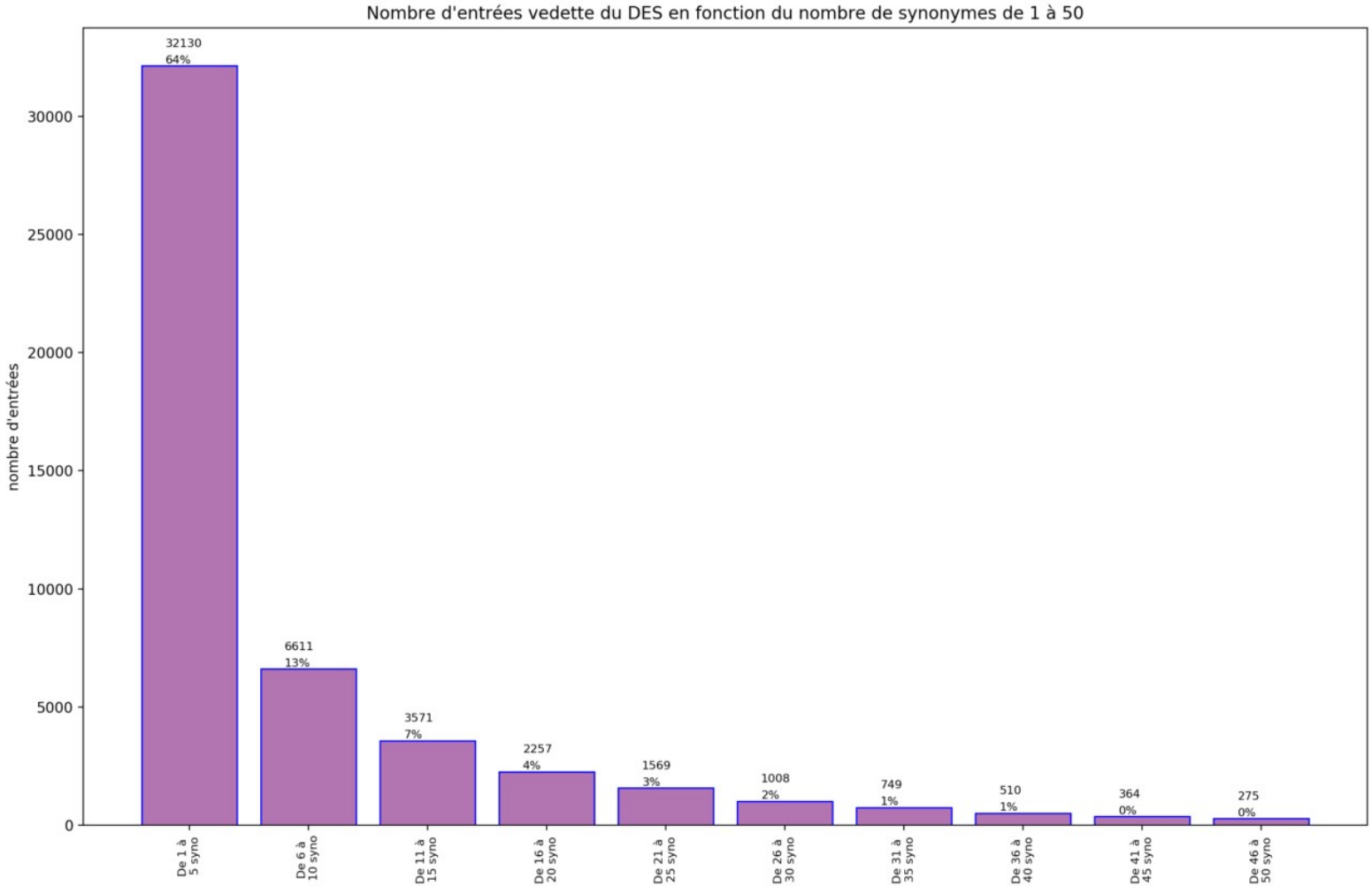


Figure 2 : Nombre entrées par nombre de synonymes (de 1 à 50)

En résumé, pour 77% des entrées, ni l'espace sémantique ni le graphe d'adjacence que nous allons développer ci-dessous, ne sont nécessaires pour comprendre les différents sens du mot (si tant est que ce soit le cas). Seul environ 23 % des entrées vont être intéressantes à visualiser avec ces outils pour repérer leur

polysémie.

⇒ Nombre de cliques

Ensuite nous pouvons nous intéresser aux nombres de cliques. Comme expliqué sur le site web ou dans cette présentation technique, une clique est un sous-ensemble le plus grand possible de sommets tous synonymes entre eux (toujours pour une vedette donnée). Si nous représentons le nombre d'entrées en fonction du nombre de cliques, comme nous l'indique la figure 3, nous déduisons quasiment la même chose que précédemment, à savoir qu'une **très grande majorité d'entrées du DES, 91% soit 47.714, possède de 1 à 20 cliques**.

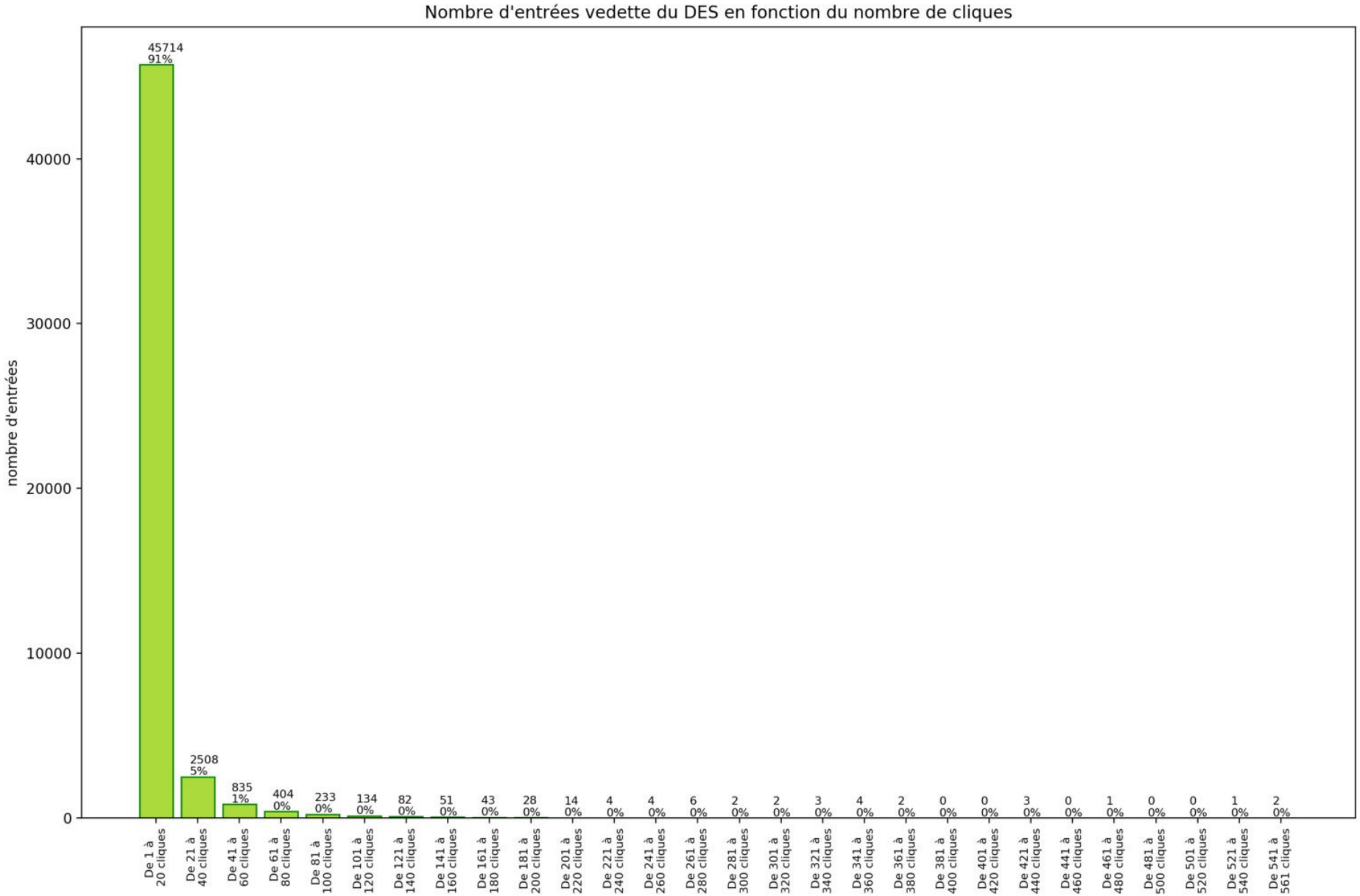


Figure 3 : Nombre d'entrées du DES en fonction du nombre de cliques (données de mars 2020)

Les entrées qui ont un très grand nombre de cliques sont sensiblement les mêmes que celles citées dans le graphique précédent : ***bon*** (561 cliques), ***prendre*** (543), ***beau*** (536), ***extraordinaire*** (469), ***faire*** (431), ***mauvais*** (430), ***fort*** (421) et ***dur*** (380).

⇒ Nombre de liens et corrélation entre ces trois indicateurs

Le nombre de liens correspond au nombre de relations synonymiques entre les synonymes d'une vedette en dehors de cette dernière.

Le graphique suivant nous apprend que **46 % des entrées (23.205) possèdent moins de 30 liens** et cela diminue fortement dans la catégorie suivante « de 31 à 60 liens » puisque nous y descendons à 6% des entrées.

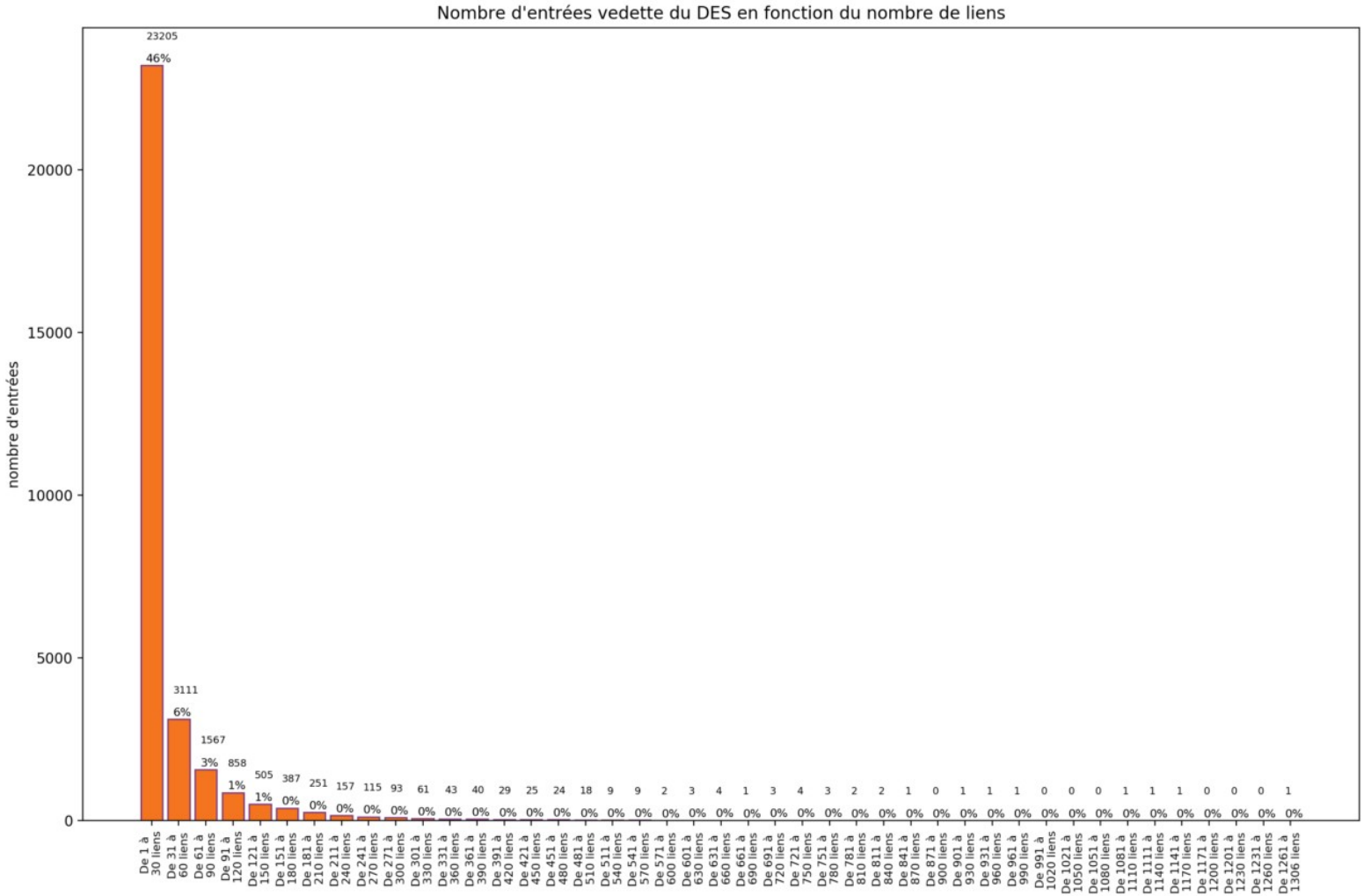


Figure 4 : Nombre d'entrées en fonction du nombre de liens (données de mars 2020)

Les trois graphiques se ressemblent. Nous en déduisons qu'il existe une corrélation entre ces trois indicateurs. Il est vrai, qu'intuitivement, nous pouvons imaginer que le nombre de cliques sera proportionnel au nombre de synonymes et que le nombre de liens sera proportionnel au nombre de synonymes et au nombre de cliques. Mais cette

proportion est-elle plutôt linéaire ou d'un autre type (polynomiale, …) ?

Pour répondre à cette question, voici ci-dessous les trois graphiques ou ces trois indicateurs sont comparés deux à deux et nous y avons joint la droite de corrélation[Note3].

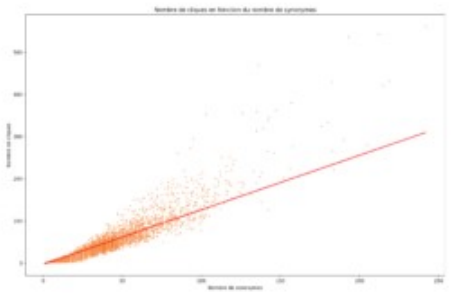


Fig 5.1 : Nombre de cliques en fonction du nombre de synonymes

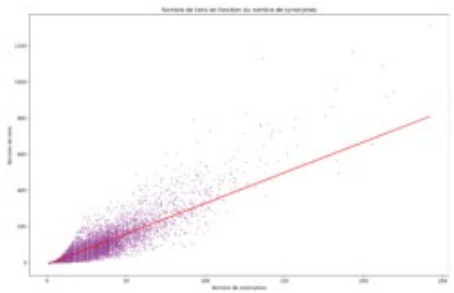


Fig 5.2 : Nombre de liens par nombre de synonymes

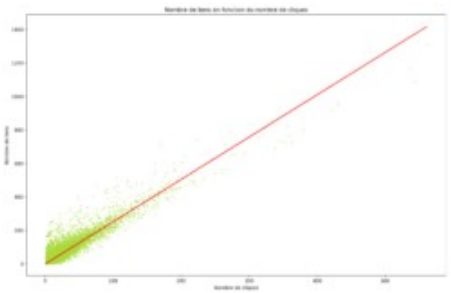


Fig 5.3 : Nombre de liens par nombre de cliques

Le coefficient de pearson[Note4] qui est la covariance entre deux indicateurs divisé par l'écart-type des deux indicateurs donne respectivement pour ces trois graphiques de gauche à droite : 0,92 (figure 5.1); 0,90 (figure 5.2) et 0,94 (figure 5.3), ce qui montre une forte corrélation linéaire entre ces indicateurs. Nous remarquons toutefois que la pente du dernier graphique (nombre de liens par nombre de cliques) est plus prononcée que celle des deux premiers graphiques. Cela s'explique par le fait que, si le nombre de cliques augmente, cela implique systématiquement que de nouveaux liens entre les sommets ont été insérés car une clique, rappelons-le est un sous-ensemble de sommets du graphe **tous reliés entre eux**. Ces deux composantes sont donc directement liées comme nous le montre par ailleurs le coefficient de Pearson (0,94). Pour les deux premiers graphiques, où la pente de la droite est moins prononcée, une augmentation du nombre de synonymes ne signifie pas que des liens nouveaux ont été créés avec les autres synonymes de la vedette mais seulement qu'un lien a été créé avec la vedette. L'ajout de synonymes a une incidence modérée et indirecte sur le nombre de cliques et le nombre de liens.

Enfin, pour finir, intéressons nous au taux de connexité autrement dit la densité d'un graphe. La densité d'un graphe est le nombre de liens réels divisé par le nombre maximum de liens possibles. La formule est donnée sur cette page wikipedia. Son application n'est pas récente puisque Jacques François a déjà réalisé des publications sur ce sujet[Note5].

Ce taux est inférieur ou égal à 1.

S'il est proche de 1, cela signifie qu'il y a beaucoup de liens entre les synonymes et logiquement moins de cliques mais elles seront plus étoffées. Parmi les taux de connexité les plus importants, nous trouvons :

Vedette	Nombre de synonymes	Nombre de cliques	Nombre de liens	Taux de connexité
impermanent	12	1	78	1
ébaubir	11	1	66	1
courbaturé	20	3	187	0,89
obsolète	18	7	150	0,87

On peut aussi remarquer à l'opposé :

extraordinaire	136	463	1262	0,13
responsable	35	30	83	0,13
étonnant	93	271	883	0,2

S'il est proche de 0, cela veut dire qu'il y a, soit peu de liens entre les synonymes, soit des sous-groupes de forte densité « noyés » dans la totalité du graphe. Dans ce dernier cas, cela correspond au différents sens du mot étudié : de nombreux liens sont présents entre certains synonymes, ce qui forme des groupes mais ces groupes sont très peu reliés les uns aux autres. Pour approfondir ces cas, le graphe d'adjacence que nous détaillons dans la section suivante va nous être très utile.

Le graphe d'adjacence

Comme nous l'avons précisé auparavant, ce graphe sera utile pour les entrées qui ont beaucoup de synonymes.

Commençons toutefois par présenter cette approche graphique avec un exemple simple : la forme « ***mousse*** » qui renvoie à deux unités lexicales MOUSSE_1(mas.) et MOUSSE_2(fém.).

Le graphe d'adjacence dans la figure 6a réalisé avec la bibliothèque igraph du langage python, nous indique que la vedette ***mousse***, représentée par le cercle central, est composée de 15 synonymes. La vedette est au centre et tout autour se déploient les quinze synonymes, tous reliés au cercle central. Certains sont également reliés entre eux selon s'ils sont synonymes ou pas. Chaque synonyme est représenté par un cercle avec un chiffre représentant le nombre de liens qu'il a. Nous remarquons que ces liens, représentés par un trait gris, sont plus ou moins épais. Cette épaisseur est proportionnelle au nombre minimum de liens à supprimer dans le graphe pour que l'un des sommets ne soit plus accessible à partir de l'autre et inversement, autrement dit c'est le nombre de liens à supprimer pour obtenir deux graphes disjoints[Note6]. Ce trait va donc traduire la force du lien : plus il sera épais, plus les deux sommets concernés sont fortement reliés, non seulement par le fait qu'ils soient synonymes l'un l'autre mais aussi par le fait que leurs synonymes à divers degrés sont également en relation entre eux.

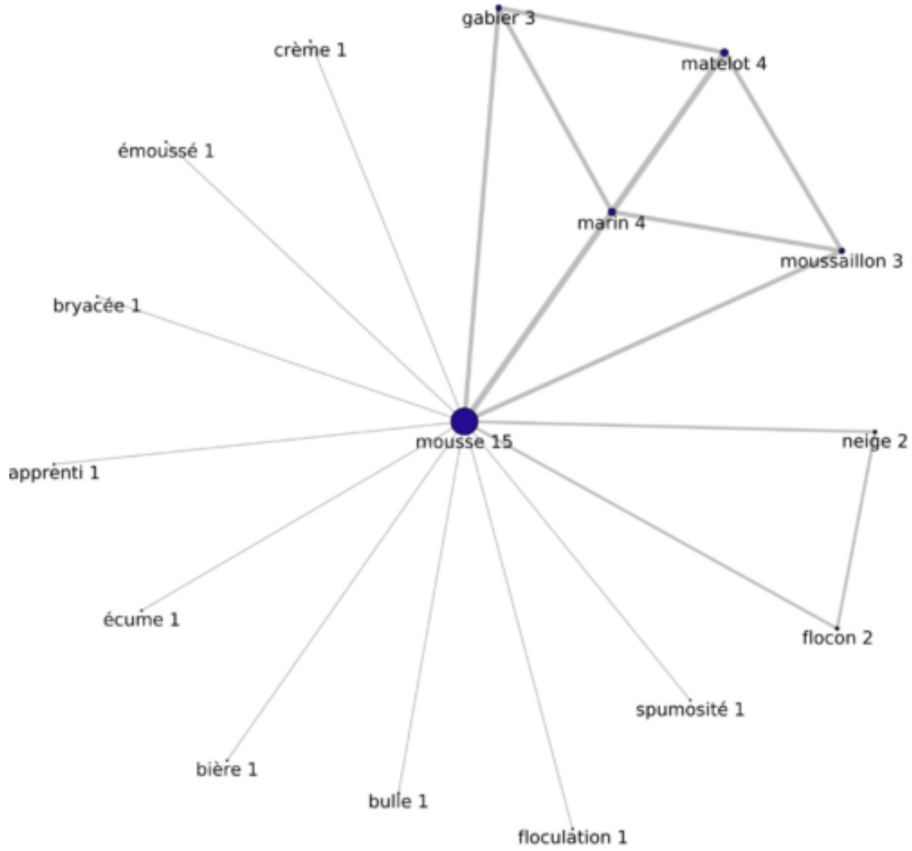


Figure 6a : le graphe d'adjacence de la vedette MOUSSE

Concrètement dans le cas de ***mousse***, le trait entre ***mousse*** et ***crème*** est fin puisqu'il suffit d'enlever uniquement ce trait pour ne plus pouvoir atteindre le sommet ***crème*** à partir du sommet ***mousse***. Entre ***mousse*** et ***neige***, il faut supprimer deux liens : celui entre ***mousse*** et ***neige*** et celui entre ***neige*** et ***flocon***. Le trait est donc un peu plus épais. Pour ***mousse*** et ***gabier***, il nous faut supprimer trois liens : ***mousse – gabier***, ***marin – gabier*** et ***matelot – gabier***. Enfin entre ***mousse*** et ***marin***, il faut supprimer quatre liens ***marin – mousse***, ***marin – gabier***, ***marin – matelot*** et ***marin – moussaillon***. De même pour ***mousse – matelot*** sachant que ce lien est superposé avec les deux liens ***mousse – marin*** et ***marin – matelot***.

Cette approche graphique est la traduction de la matrice d'adjacence, MatAdj, figure 6b, qui contient en ligne comme en colonne la vedette et ses synonymes avec $MatAdj_{(i,j)}$ égal à 1 si le terme sur la ligne i est synonyme du terme sur la colonne j, égal à 0 dans le cas contraire. En voici une représentation :

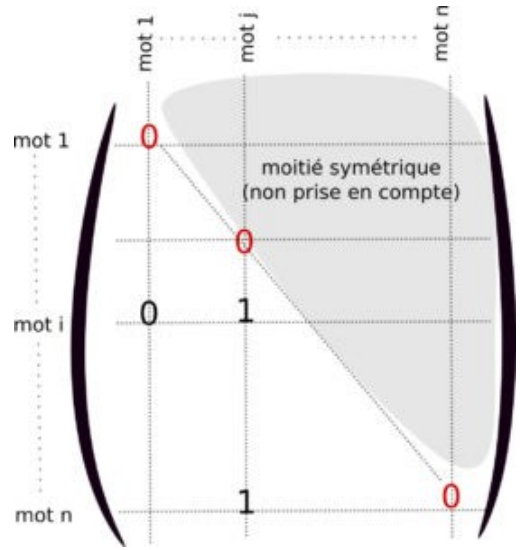


Fig 6b : Matrice d'adjacence

Reprenons maintenant, notre exemple avec ***impermanent*** qui a une densité de 1. Comment cela se traduit-il dans le graphe d'adjacence ? La réponse apparait évidente lorsque nous voyons ce graphe ci-dessous : tous les sommets sont reliés entre eux. Autrement dit, tous les liens possibles sont réellement effectifs.

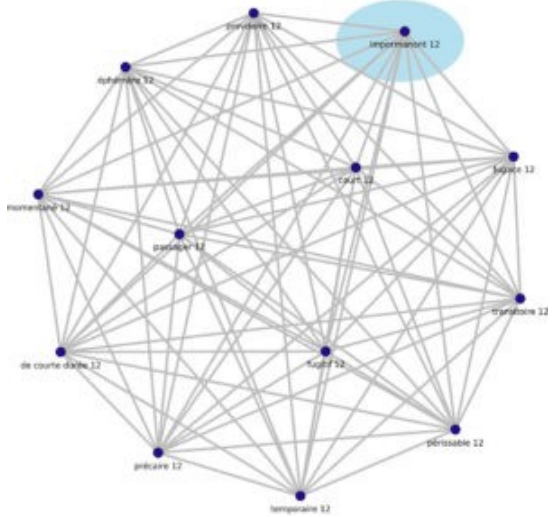


Figure 7 : Graphe d'adjacence de IMPERMANENT

Nous pouvons aussi revenir sur la vedette ***courbaturé*** qui a un taux de connexité de 0,89. Dans ce cas aussi nous retrouvons un graphe très fortement connecté avec toutefois avec le synonyme ***ankylosé*** qui n'est connecté qu'à la vedette et à un de ses synonyme ***courbatu***. Cette position un peu particulière explique ce taux qui serait beaucoup plus proche de 1 sans cela.

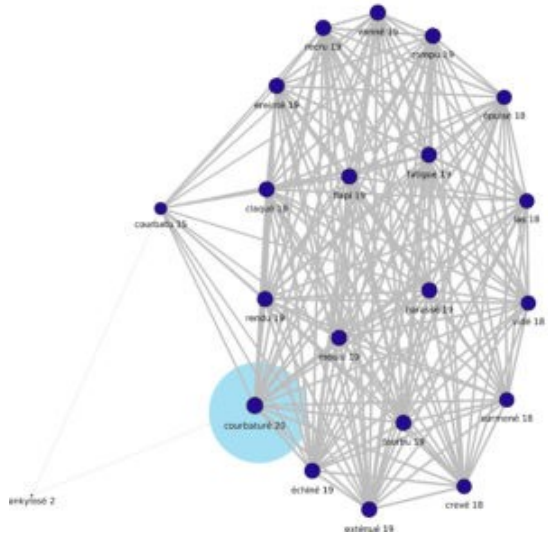
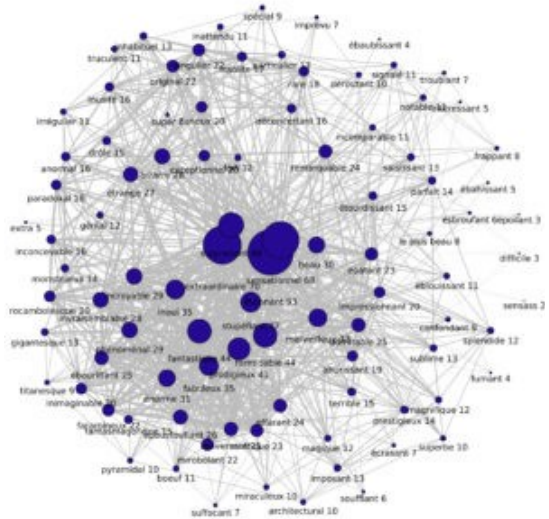


Figure 8 : Graphe d'adjacence de COURBATURE

Nous allons maintenant nous intéresser aux mots qui possèdent un grand nombre de synonymes comme ***étonnant*** qui a une densité assez faible de 0,2. Son graphe est représenté par la figure 9.1.



La Figure 9.1 : Graphe d'adjacence de ÉTONNANT

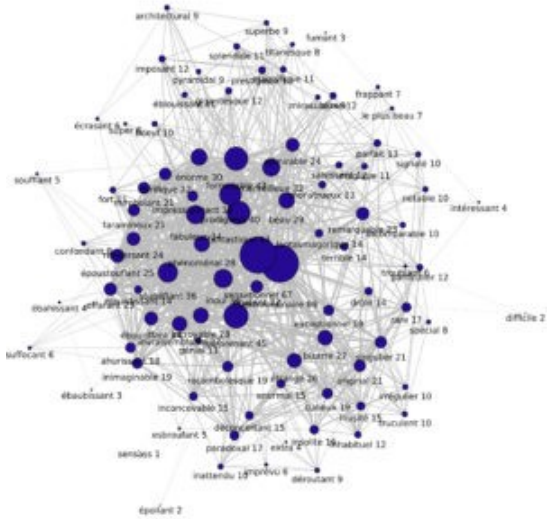


Figure 9.2 : Graphe d'adjacence de ÉTONNANT sans la vedette

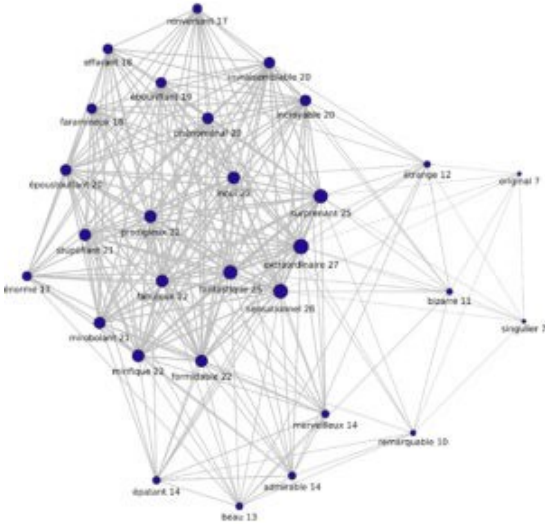


Figure 9.3 : Graphe d'adjacence de ÉTONNANT sans la vedette ni les

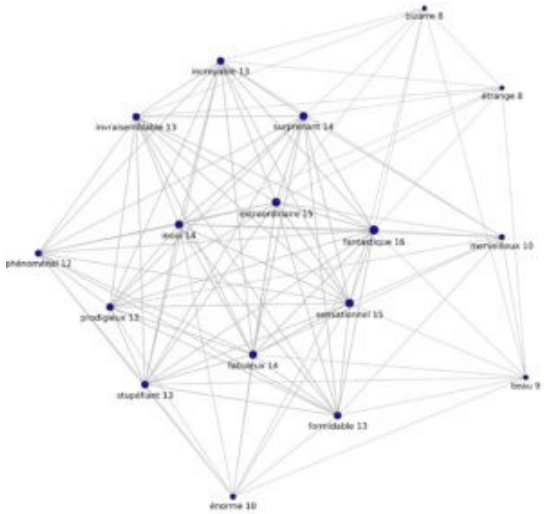


Figure 9.4 : Graphe d'adjacence de ÉTONNANT sans la vedette ni les

synonymes qui ont moins de 20 liens

synonymes qui ont moins de 25 liens

A première vue, il peut paraitre curieux d'avoir une densité si faible alors que nous voyons un graphe assez touffu. En réalité, cette densité est fortement diminuée par les synonymes à la périphérie qui sont assez peu connectés aux autres au centre. Pour obtenir un graphe plus clair, nous pouvons enlever la vedette. En effet, celle-ci étant reliée à tous ses synonymes, sa suppression va s'accompagner de la suppression de tous les liens qu'elle a. Nous obtenons le graphe de la figure 9.2. Malheureusement, ce n'est pas suffisant. Une autre solution ensuite est de supprimer les sommets qui ont moins de N liens, en augmentant progressivement la valeur de N. Ces sommets sont comme nous pouvons le deviner ceux qui seront en périphérie. La figure 9.3 nous donne un résultat en supprimant tous les sommets qui ont moins de 20 liens. Le graphe est ainsi plus facile à exploiter : nous voyons ainsi un amas de synonymes très reliés entre eux : ***prodigieux, inouï, extraordinaire,surprenant, sensationnel, fantastique*** … et d'autres synonymes sur la partie droite du graphique un peu plus dispersés : ***merveilleux, bizarre, étrange épatant***. Si nous allons plus loin et enlevons tous les sommets qui ont jusqu'à 25 liens, le graphique obtenu en figure 9.4 nous donne un graphe avec seulement 17 sommets qui se concentre uniquement sur les synonymes les plus nombreux à savoir ***surprenant, sensationnel, fantastique*** alors que ***bizarre et étrange*** sont quelque peu marginalisés. Dans ces quatre graphiques, la densité augmente : nous passons de la gauche vers la droite de 0,2 à 0,33 puis 0,65 et enfin à 0,77.

Augmenter la densité peut être un critère pour permettre d'avoir un graphe plus clair qui reflète surtout la polysémie de la vedette, mais elle doit être utilisée à bon escient. Si nous prenons l'exemple de ***responsable***, dont la densité du graphe initial présenté dans la figure 10.1 est de 0,131, nous passons à une densité très légèrement supérieure (0,132) avec un graphe sans la vedette puis en supprimant uniquement les sommets ayant un seul lien (figure 10.2). Ce graphe est suffisamment explicite pour en déduire les différents sens de responsable : ***représentant***, ***tuteur-trice*** puis ***condamnable***, ***coupable*** et enfin ***directeur-trice***, ***dirigeant-e***.

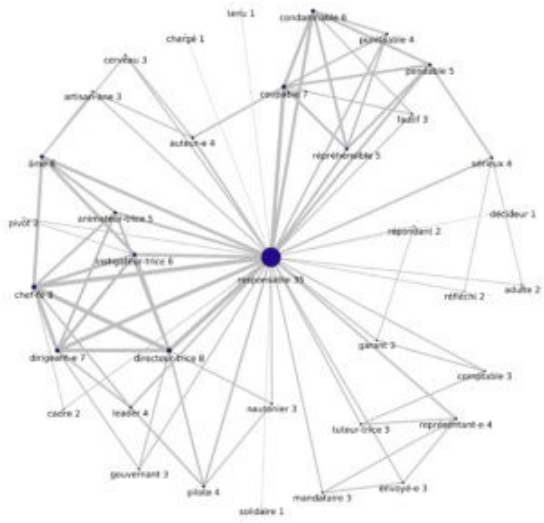


Figure 10.1 : Graphe d'adjacence de RESPONSABLE

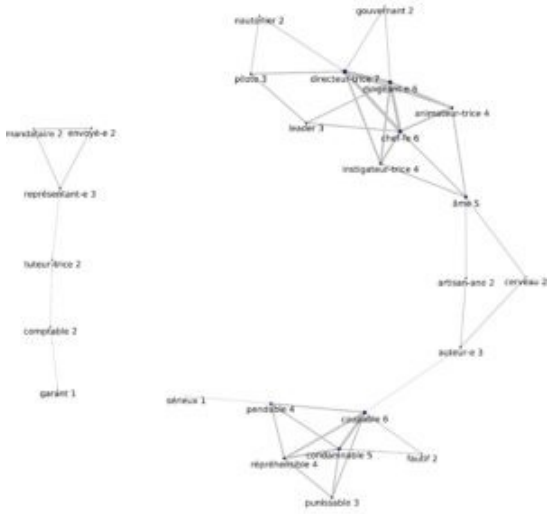


Figure 10.2 : Graphe d'adjacence de RESPONSABLE sans la vedette et les sommets de moins de 1 lien

Conclusion

Le fait d'enlever les sommets qui ont le moins de liens avec les autres sommets permet d'éclaircir le graphe. Comme la méthode est générale et s'applique à tous les sommets, si nous enlevons un trop grand nombre de sommets ayant moins de N liens, nous n'allons garder que les sommets les plus connectés représentant le sens le plus courant de la vedette. Les autres sens tout aussi importants qui sont représentés par des sommets qui ont moins de liens disparaitront. Un autre exemple significatif étudié lors de la lettre d'actualité n°6 concerne le verbe ***gagner***. Ce verbe a de nombreux sens comme par exemple ***s'étendre, se propager, se répandre, …***. Ce sens apparait dans le graphe de base, et dans celui ou la vedette est absente ainsi que les synonymes de moins de cinq liens. Mais il disparait lorsque nous enlevons les sommets qui ont jusqu'à 10 liens (voir les figures 11.1, 11.2 et 11.3 ci-dessous)

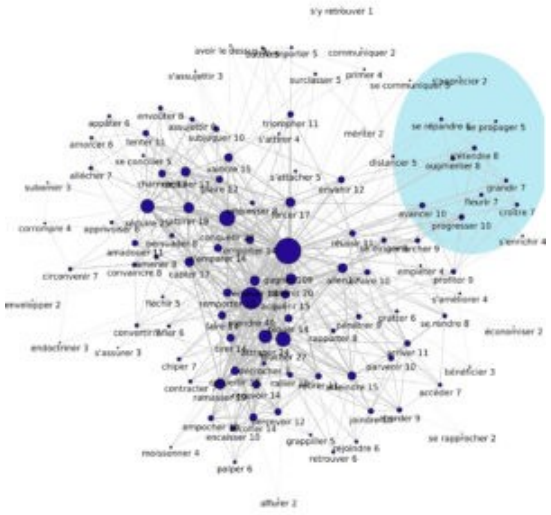


Figure 11.1 : Graphe d'adjacence de GAGNER

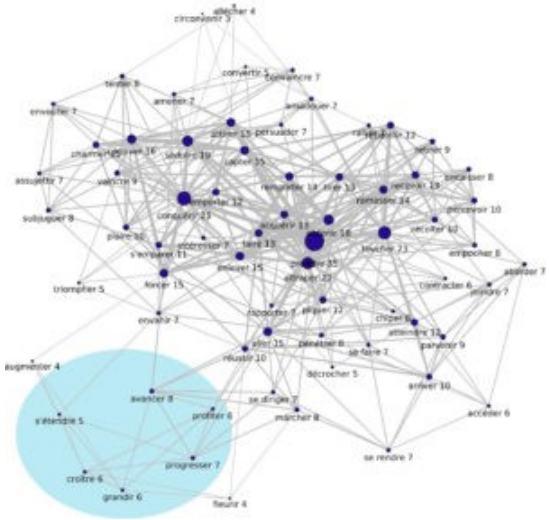


Figure 11.2 : Graphe d'adjacence de GAGNER sans la vedette ni les sommets ayant mois de 5 liens

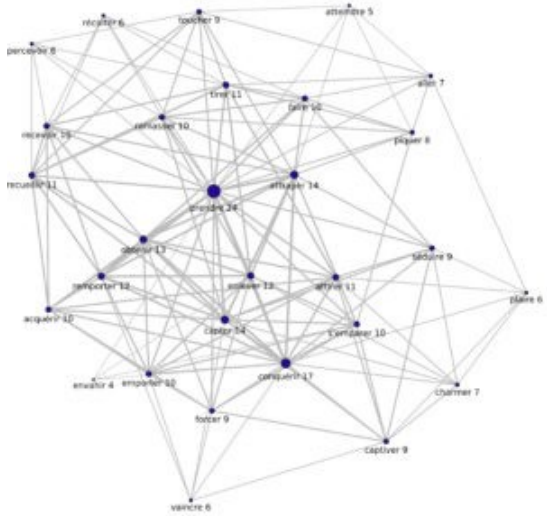


Figure 11.3 : Graphe d'adjacence de GAGNER sans la vedette ni les sommets ayant mois de 10 liens

La représentation d'une vedette avec son graphe d'adjacence est un moyen rapide d'obtenir des réponses aux questions suivantes : le graphe est-il clairsemé ou au contraire touffu ? Y a t-il beaucoup de synonymes en périphérie ? Les sommets sont-ils en majorité reliés à la vedette uniquement ? Voit-on des groupes de mots ? Ces groupes sont-ils reliés entre eux ou au contraire distincts ? … De plus, les options que nous pouvons utiliser : affichage ou pas de la vedette, effacement des sommets ayant moins de N liens vont faciliter les réponses.

Cette méthode va aussi nous permettre de nous interroger sur les sommets ayant peu de synonymes : cela est-il dû à des liens synonymiques non saisis ? ou bien correspondent-ils à une facette de la polysémie de la vedette qu'il faudrait garder ? La suppression des liens périphériques doit donc s'accompagner d'autres outils comme des méthodes de partitionnement de données (ou clusterisation) ou bien des outils proposés par Bruno Gaume dans sa structure de « petit-monde », terme employé pour les graphes sémantiques lexicaux qui se caractérisent par des distances moyennes très petites relativement à la taille du graphe et des structures locales très denses[Note7] .

Merci à Michel Morel et Jacques François pour leur relecture

Notes

Note1

L'insertion des liens synonymiques est manuelle. Elle n'est donc pas exempte d'erreurs de saisie et de quelques liens synonymiques incohérents, même si la démarche utilisée les minimise au maximum. Une interface de proposition est donc à votre disposition pour nous permettre de corriger et d'améliorer le DES. Elles sont traitées le mois suivant votre dépôt et mises en ligne, après vérification sur le site le mois d'après.

Retour à la référence de la note

Note2

Les liens vers les deux vidéos de remerciements :

http://crisco.unicaen.fr/dictionnaire-electronique-des-synonymes/actualites-des/remerciements-des-utilisateurs-du-des-920615.kjsp?RH=1530619460865

et

http://crisco.unicaen.fr/dictionnaire-electronique-des-synonymes/actualites-des/remerciements-des-utilisateurs-du-des-2nde-video–947671.kjsp?RH=1530619460865

Retour à la référence de la note

Note3

La fonction utilisée en python est scipy.stats.linregress

Retour à la référence de la note

Note4

Pour une définition plus complète du coefficient de Pearson, se référer à la page sur wikipedia

La fonction utilisée en python est scipy.stats.pearsonr

Retour à la référence de la note

Notes5

Le cahier du CRISCO n°20 (2005) et FRANCOIS Jacques (2007), *Pour une cartographie de la polysémie verbale* (2007), Peters Louvain, Chapitre II

Retour à la référence de la note

Note6

Nous utilisons la fonction edge_connectivity() de la librairie igraph : https://igraph.org/r/doc/edge_connectivity.html

Retour à la référence de la note

Note7

Voir à ce sujet ces deux publications : Quand les mots s'organisent en réseaux et Balades aléatoires dans les Petits Mondes Lexicaux.

Retour à la référence de la note