Perrine MAUREL – Mars 2022

Rapport : Devoir final de R

École Nationale des Chartes & EPHE

UE : Introduction à l’analyse et visualisation de données avec le logiciel libre R

Sous la direction de Daniel Stockholm

Perrine MAUREL – Mars 2022

I. Présentation du corpus et de sa problématique

Le corpus choisi est un site Internet d'écriture de l'imaginaire et de traduction communautaire, *La Fondation SCP*. Il regroupe plus de 45 000 textes écrits par environ 7500 auteurs à travers quinze langues différentes. La taille du corpus premier étant assez élevée, j'ai décidé d'effectuer une préselection en ne me concentrant que sur les textes de la branche francophone, puis en restreignant mon corpus aux formats des 'contes' et des 'rapports' qui sont les deux formats prédominants sur le site.

Tous les textes se déroulent sensiblement dans un même univers qui s'étoffe petit à petit avec les contributions des différents auteurs. On observe ainsi l'apparition de nombreux néologismes locaux, propres à cet univers, dont l'utilisation varie dans le temps, dans l'espace linguistique et dans l'espace social du site.

Par exemple :

amnésique : médicament permettant d'effacer la mémoire (autre orthographe : *amnésiant*)
gendastre : membre de la Gendastrerie, un corps de la police nationale française

L'étude portait donc sur l'existence de ces mots et leur utilisation à travers les différents textes. J'ai aussi accumulé des informations portant sur les métadonnées des textes afin d'ouvrir nos horizons à d'éventuelles analyses supplémentaires, qui n'ont pas encore été mises en place.

—

La difficulté résidait surtout dans l'interaction Web qui ne m'est pas encore familière, tout particulièrement puisque les pages HTML faisaient des appels au serveur pour récupérer certaines métadonnées et qu'il me fallait donc créer une session et reproduire cet appel pour obtenir mes données. J'ai donc demandé de l'aide à Corentin POUPRY, lequel avait écrit un programme à cette fin en Python, disponible ici sur son GitHUB, pour m'aider à comprendre le principe de son code et pour le retranscrire en R. Je l'en remercie grandement.

Du fait de la taille du corpus, certains algorithmes prennent un certain temps pour s'exécuter. J'ai ajouté à mes dataframes une colonne indiquant le temps pour chaque requête, et à certaines de mes fonctions une option pour calculer le temps exact qu'elles mettent à s'exécuter, afin de pouvoir quantifier exactement le temps de calcul de mes algorithmes et de comprendre d'où venaient les ralentissements. Idéalement, je pourrai améliorer le temps de production des données en revisitant mon code dans le futur.

II. Méthodes

1. Construction du corpus

J'ai donc créé un dataframe `doudou` de 1272 objets grâce à la fonction `recherche_liste_tag_fr()`, chaque ligne correspondant aux métadonnées du texte ainsi qu'au texte lui-même. Un autre dataframe `meta_doudou` contient ce même nombre d'objet : c'est que ce second objet récupère les métadonnées de la page HTML grâce à une requête ajax, ce qui demande un petit peu plus de travail et l'utilisation de la bibliothèque `crul`, là où `doudou` récupère les métadonnées à partir du module de crédit ajouté manuellement par les utilisateurs du site.

Une fois ces données assemblées, j'ai décidé dans le cadre du rapport de me concentrer sur les néologismes et j'ai donc créé une liste non-exhaustive de quinze exemples choisis. Certains objets de cette liste correspondent au mot en entier ; d'autres, comme `'cinquis'`, correspondent à une racine étymologique permettant de trouver toutes les déclinaisons du mot.

2. Étude du corpus

J'ai ensuite observé l'existence de ces néologismes dans mon corpus sur la base de deux critères :

1. La présence, c'est-à-dire le nombre d'occurrence dans le corpus
2. La fréquence, c'est-à-dire le pourcentage que représentent ces néologismes dans le corpus

Ces deux notions étant intrinsèquement liées, j'ai fini par ne plus me concentrer que sur la fréquence.

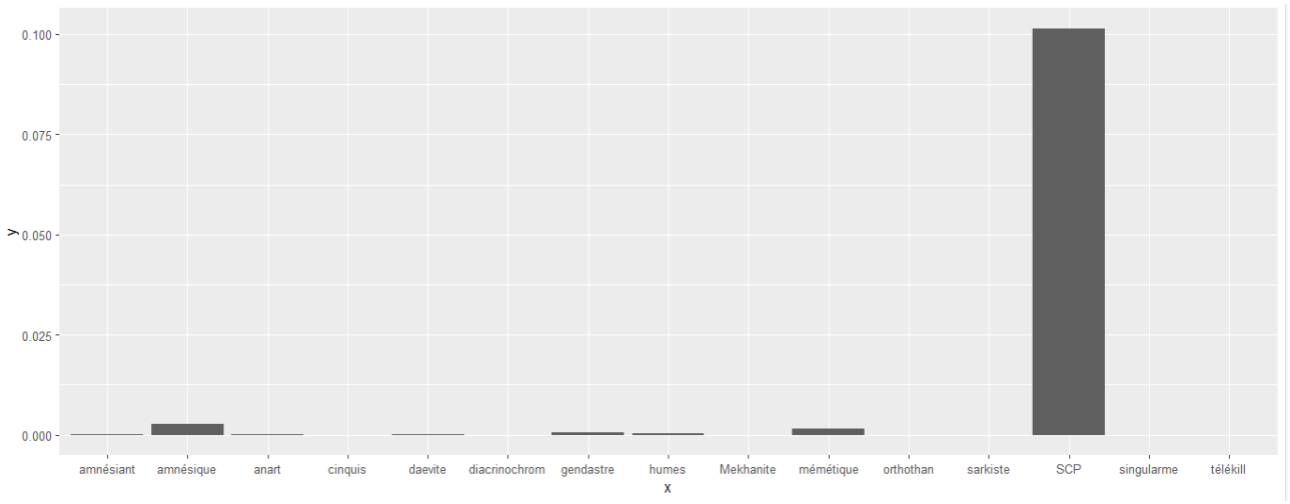

Figure 1: Graphe `df`*, néologismes en abscisses et fréquence*10 en ordonnées*

Lors de mes premiers tests, le graphe en barres `df` montrait que le néologisme `'SCP'` était largement plus utilisé que les autres. J'ai donc voulu calculer la variance et l'écart-type afin de confirmer mon observation de manière empirique, et j'ai trouvé une variance de près de `20` pour la présence, ce qui est non-négligeable.

Pour expliquer ce phénomène, j'ai émis l'hypothèse que le néologisme principal 'SCP' l'était car il était très fortement utilisé dans les rapports, là où les autres néologismes étaient utilisés de manière plus égale et ponctuelle à travers les différents formats. Pour tester cette hypothèse, j'ai décidé de calculer la fréquence des néologismes dans les contes et celle dans les rapports, puis de compiler ces informations dans un graphique avec l'application de la méthode de réduction directionnelle ACP.

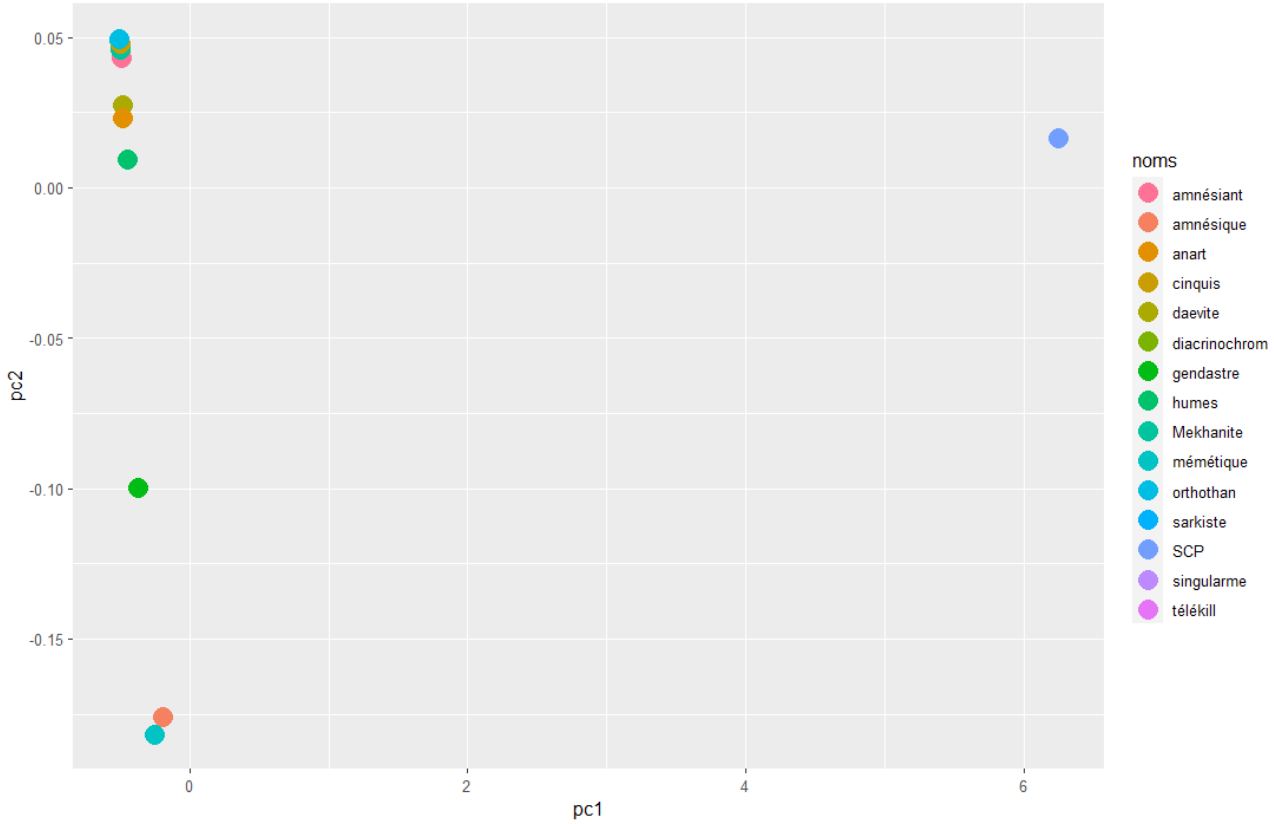

Figure 2: Graphe `df_stat`*, compare la fréquence des néologismes dans le corpus entier, dans les contes et dans les rapports*

3. Résultats et coda

L'hypothèse concernant le néologisme 'SCP' semble être validée. Il serait alors intéressant de l'exclure pour observer le comportement des autres néologismes plus rares.

Pour continuer sur cette lancée, je voudrais examiner le parcours chronologique et évolutif des termes en observant les auteurs et les textes qui les auraient repris, pour obtenir une modélisation visuelle en graphe. Il serait alors possible d'examiner les caractéristiques grammaticales et contextuelles de ces néologismes afin de comprendre ce qui fait la popularité de certains, et pas d'autres.

J'aimerais également me concentrer sur d'autres champs d'études, par exemple en confrontant les données de `meta_doudou` et de `doudou` afin d'obtenir des statistiques sur l'efficacité des utilisateurs lorsqu'ils retranscrivent et complètent manuellement les métadonnées d'un texte.