Projet de R : analyse des évolutions grammaticales de deux corpus en langue corse

Vincent Sarbach-Pulicani

15 mars 2022

1 Introduction

Depuis les années 1980, il existe en Corse une volonté de structurer et d'étudier les évolutions de l'emploi de la langue corse . On peut notamment mentionner les travaux de la linguiste Marie-José Dalbera-Stefanaggi avec son *Nouvel atlas linguistique et ethnographique de la Corse* dont le premier volume paraît en 1995 [1]. Dans les rééditions de cette œuvre majeure dans les années 2000, l'autrice incorpore notamment ses travaux sur la création d'une Banque de Données Langue Corse (BDLC). Il s'agit là de la première initiative de lemmatisation de la langue corse dans l'espace et le temps. L'apport des humanités numériques dans cette problématique est assez neuf. Les réflexions des chercheurs sur l'outillage des langues régionales en utilisant le TALN (Traitement Automatique du Langage Naturel) se sont vraiment accélérées à partir de la deuxième moitié de la décennie 2010. Face au manque d'outils pour traiter et analyser la langue corse [2], la moindre étude se révèle pertinente pour l'avancée de la recherche et la préservation du patrimoine linguistique insulaire. L'objectif de notre travail s'inscrit donc dans cette logique. L'objectif est d'évaluer les évolutions grammaticales de deux corpus de textes originaires d'époques différentes.

Le premier est composé de plusieurs textes variés publiés dans les années 1920 par la *Stamparia di a Muvra*, organe de publication des « muvristes ». Mouvement autonomiste corse de l'entre-deux-guerres incarné par la revue *A Muvra*, ces-derniers étaient les producteurs principaux de textes dialectaux. En effet, la langue corse est dite « langue du XX^{e} siècle » et les règles grammaticales qui structurent le langage écrit n'en sont qu'à leurs balbutiements.
Le second corpus est une traduction contemporaine du *Comte de Monte Cristo* disponible sur le site de la revue dialectale *Tempi Rivista*. Les textes sont variés et abordent des thèmes et styles différents, mais notre travail se concentrera essentiellement sur les mots-outils donc cela n'est, en principe, pas un problème pour nous.

À terme, nous voulons savoir sous quelles formes prennent les évolutions grammaticales les plus basiques en 100 ans d'écriture de la langue corse.

1. Dalbera-Stefanaggi Marie-José, *Nouvel Atlas Linguistique et Ethnographique de la Corse*, Paris, éditions du CNRS, 1995.
2. Kevers Laurent, Gueniot Florian, Tognotti A. Ghjacumina, Retali-Medori Stella, « Outiller une langue peu dotée grâce au TALN : l'exemple du corse et BDLC », *26e Conférence sur le Traitement Automatique des Langues Naturelles*, Toulouse, 2019, p. 371 à 380.

2 Méthodologie et limites

Dans un premier temps, il était nécessaire de nettoyer correctement nos textes. Pour cela, j'ai créé une fonction `clean_text` qui, à l'aide d'expressions régulières, se débarrassait de la ponctuation et de la casse. Cette fonction sert également à normaliser les textes de deux manières :

- Une normalisation de l'encodage, les textes provenant pour certains d'eScriptorium, l'unicode n'est pas forcément respecté, particulièrement au niveau des accents, faussant ainsi les résultats.
- une normalisation de l'écriture, les textes anciens (OLD) ayant une graphie différente des textes contemporains (NEW). Par exemple, le verbe être à la 3^{e} personne du singulier s'écrit soit « è » ou soit « hè ».

Les 8 textes sont ensuite divisés en deux dossiers pour former les deux corpus différents, qui seront par la suite importer dans RStudio. Séparer les deux corpus est très important car nous sommes dans le cadre d'une étude comparative : si le script peut en effet être plus long et moins écologique, cela reste la manière la plus efficace de comparer deux *dataset* entre eux.
Dans la première phase de notre analyse, j'ai créé une *dataframe* qui contient le nombre de caractères totaux de chacun des corpus puis le nombre de mots une fois la tokénisation effectuée. J'ai ensuite fait le choix de calculer le nombre de caractères moyens par mots, élément intéressant pour déterminer s'il y a un phénomène de littérarisation de la langue corse ou non. nous observerons les résultats obtenus dans notre troisième partie.
Une fois ce premier tableau effectué, qui compare des données relativement basiques des deux corpus, il a fallu calculer les fréquences de mots, ce qu'on appelle les MFW (*Most Frequent Words*). C'est à cette étape que la constitution de deux corpus différents prend tout son sens. J'ai donc créé deux objets *dataframes* comportant les 15 MFW de chaque corpus avec leur fréquence absolue et leur taux en pourcentage donnant en sortie deux histogrammes.
Enfin, dans une dernière étape, j'ai effectué une réduction de dimension du calcul de la fréquence des mots afin de produire une matrice de correspondance du taux de fréquence de chacun desdits mots dans les textes du corpus. Cela permet ensuite de visualiser ces résultats à travers une *heatmap* classifiée par des dendrogrammes permettant de repérer des groupes de mots plus ou moins utiliser dans les corpus. La métrique utilisée pour réaliser ces dendrogramme est la distance euclidienne, la métrique de base proposée par RStudio.

Durant cette étude, j'ai fait face à quelques problèmes et découvert les limites d'une telle méthodologie. D'abord, j'ai eu des difficultés à effectuer la réduction dimensionnelle, les matrices en sortie n'étant pas totalement satisfaisantes selon moi, même si la projection de données me semble relativement cohérente.
Ensuite, le corpus a présenté deux problèmes qui tendent à légèrement fausser les résultats : d'abord vis-à-vis des différences de graphies évoquées précédemment, comme la négation « ùn » où l'accentuation n'est pas présente dans les textes anciens, causant une ambiguïté par homonymie syntaxique avec le déterminant cardinal « un » ; d'où la limite d'avoir un texte non lemmatisé car le contexte ne peut être pris en compte pour marquer la différence. Puis d'un point de vue sémantique, les textes anciens sont variés, allant de la monographie à la poésie, alors que les textes contemporains sont une traduction. L'usage fait des mots-vides n'est pas forcément le même et donc leur quantité peut varier suivant le type de texte.
Enfin, certains textes provenant d'un processus d'océrisation, il peut y avoir des déchets de transcriptions provoquant du bruit dans mes corpus, d'où le besoin de nettoyer une deuxième fois le texte. Ceci étant dit, cela ne doit pas avoir modifié les résultats finaux.

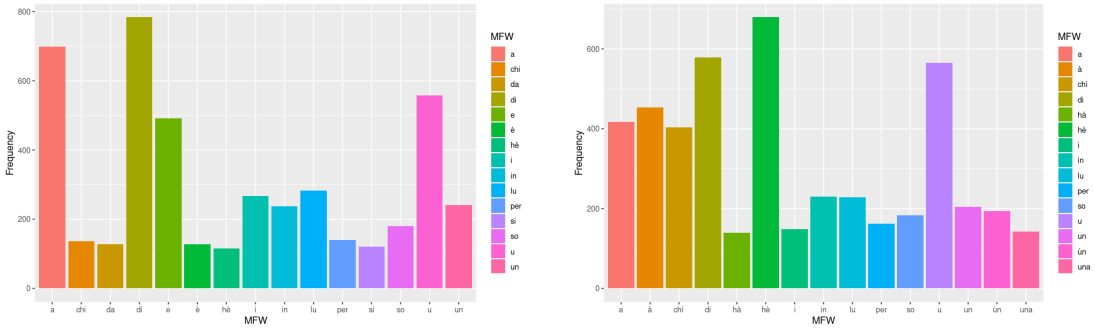
3 Interprétation des résultats

La première phase du script nous a permis d'obtenir ce tableau comparatif des caractéristiques globales des deux corpus :

TABLE 1 – Tableau représentant les caractéristiques des deux corpus

Corpus	Nb de caractères	Nb de mots	Caractères par mots
Old	84634	14894	5.682422
New	77733	14978	5.189812

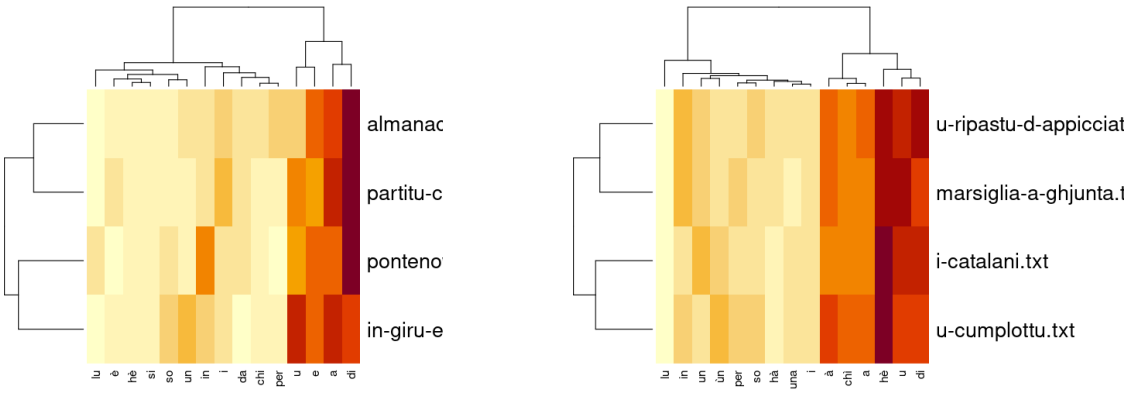
La deuxième phase de notre étude nous a permis de visualiser ces deux histogrammes :



(a) Corpus OLD (b) Corpus NEW

FIGURE 1 – Histogramme des MFW-15

Enfin, la troisième phase de notre étude est visualisée par ces deux *heatmap* :



(a) Corpus OLD (b) Corpus NEW

FIGURE 2 – *Heatmap* des MFW-15

Le tableau 1 permet de nous rendre compte que, pour un texte de taille relativement équivalente, les mots sont en moyenne 10% plus long dans les textes anciens que dans les textes

contemporains. Cela peut s'expliquer peut-être par le genre de textes, comme nous avons vu précédemment, ou alors dans la volonté de littérarisation d'une langue principalement orale dans les années 1920, en cherchant à complexifier le langage pour la rendre crédible. Ce constat est hypothétique évidemment, mais correspond néanmoins à une forme de réalité du terrain avec la floraison de nombreux romans et nouvelles en langue corse à l'entre-deux-guerres alors que le corse contemporain s'écrit davantage dans la vie quotidienne, poussant les auteurs de nos jours à, peut-être, adopter des pratiques orthographiques et grammaticales moins complexes.
Selon les figures 1, le mot le plus utilisé dans le corpus OLD est l'article « di » alors que la forme conjuguée du verbe *esse* « hè » est le terme le plus employé dans le corpus NEW. Les différences entre les deux corpus ne se font pas uniquement au niveau des mots-outils en eux-mêmes. Les graphiques nous montrent également une plus grande homogénéité dans l'utilisation des différents mots-outils de la langue corse dans les textes contemporains. À l'inverse, les textes anciens utilisent essentiellement 4 mots : « a », « di », « u », « e ». Encore une fois, faisons attentions aux ambiguïtés induisent par une graphie plus simpliste à l'époque de l'entre-deux-guerres.
Les figures 2 appuient encore plus ce résultat, les *heatmap* permettent de mettre en valeur plus clairement deux groupes principaux d'emploi des *stopwords* dans chaque corpus grâce à la classification ascendante hiérarchique des mots-outils.

Pour conclure, une telle étude est prometteuse car elle permet de réellement mettre en évidence des évolutions grammaticale dans la langue corse. La structuration d'un jeune langage principalement oral prend du temps et nécessite de la pratique par de nombreux acteurs pour qu'elle prenne en consistance. Les auteurs contemporains semblent insister sur l'utilisation du verbe être et du pronom relatif « chì », ce-dernier étant peu utilisé dans par les auteurs autonomistes. La seule concordance se fait au niveau des pronoms définis mais ceci étant la base des groupes nominaux, il y a là une forme de logique dans la perpétuation de ces mots-outils précis.
Les perspectives de recherche et d'amélioration sont nombreuses. Dans l'idéal, il faudrait se munir de textes plus propres et semblables, mais cela se frotte au problème de disponibilités des *dataset* de textes en langues corses numérisés, le corse étant un langage dit « peu doté ». Néanmoins, l'espoir est permis d'avoir de meilleurs jeu de données afin de poursuivre les analyses en allant plus loin : lemmatisation, nettoyage et normalisation efficace des textes, ... Autant d'enjeux pouvant aboutir à des études plus approfondies sur les évolutions de la langue corse, qui ne nous a pas livrée tous ses secrets.