UVSQ université PARIS-SACLAY
Institut d'études culturelles et internationales
CAMPUS DE SAINT-QUENTIN-EN-YVELINES

Livrable d'Humanités numériques : Voyant Tools et l'analyse de données textuelles

Elodie Goupilleau

I) État de l'art

Les Humanités Numériques renvoient au traitement des données numériques en sciences humaines, et ce par l'utilisation d'une multitude de méthodes et de logiciels mis à disposition des chercheurs en fonction des sources mobilisées et des résultats souhaités.
Dans ce livrable, nous nous concentrerons sur une méthode en particulier : l'analyse de données textuelles à travers l'outil Voyant Tools. Avant de dresser une description de ce logiciel, qu'est-ce que l'analyse de données textuelles ? Le *Text Mining* ou la textométrie, est une méthode issue des deux champs de recherche que sont la statistique lexicale, qui permet de décrire le vocabulaire des textes, et l'analyse multidimensionnelle lexicale qui « a mis au point un ensemble de méthodes dédiées au traitement statistique des tableaux de données dont l'analyse factorielle des correspondances »[1]. La textométrie est donc l'analyse de données textuelles est une méthode qui permet de traiter un texte ou un corpus textuel en s'appuyant sur des approches qualitatives et quantitatives. Alors que la première se concentre par exemple sur les correspondances entre les textes ; la seconde approche se traduit par une attention portée aux spécificités du textes ainsi qu'aux concordances[2]. Les spécificités permettent de mettre en perspective les termes qui sont « anormalement fréquents » dans un texte et leur présence dans le reste du texte. Quant aux concordances, il s'agit d'identifier, à partir des informations dont on dispose sur un terme, les motifs de celui-ci.[3] À présent, il convient de présenter l'outil dont il est question ici.

Créé en 2016 par les chercheurs canadiens Stéfan Sinclair (Université McGill)[4] et Geoffrey Rockwell (Université de l'Alberta), Voyant Tools[5] est un logiciel open source de Text Mining, accessible en ligne depuis 2008 puis en 2015 pour Voyant 2.0. Il a pour but de faciliter la lecture et l'analyse d'un corpus textuel.

Si cette méthode est employée progressivement dans de nombreuses disciplines, il faut également souligner la multitude de logiciels qui existent pour automatiser l'analyse des textes, signe de l'interêt grandissant pour les humanités numériques. Parmi eux, afin de replacer notre logiciel dans une production plus large, et ce de manière non exhaustive, nous pouvons en citer quelques-uns qui semblent présenter des fonctionnalités assez similaires à **Voyant Tools**. **IRaMuTeQ** qui, même s'il fonctionne avec le langage Python, permet également de visualiser les données sous forme de listes ou encore de graphes arborescents présentés dans l'article de Bénédicte Pincemin[6] comme les spécialités de cet outil parmi diverses autres visualisations possibles. Dans ce même article, il est également question du logiciel **Hyperbase Web Edition**. Créé en 2010, il découle d'une première version d'**Hyperbase** de 1989. Tout comme dans Voyant Tools, les calculs de spécificités ou encore l'étude des concordances sont possibles avec cette

---

[1] Damon Mayaffre, Bénédicte Pincemin, Céline Poudat, « Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse des discours », Langue française, Armand Colin, 2019, Les outils informatiques au service des linguistes, pp.101-115. hal-02419199, p. 2.

[2] Bénédicte Pincemin, *Sept logiciels de textométrie,* 2018. halshs-01843695, p.1.

[3] « Explorer, mesurer, contextualiser », *op.cit,* p. 6.

[4] Ce dernier avait déjà développé dans les années 2000 un autre logiciel dans le cadre de recherches universitaires : l'Oulipo.

[5] Voyant Tools s'inscrit plus largement dans un projet que nous pouvons retrouver dans le livre suivant écrit par les deux créateurs du logiciel : Geoffrey Rockwell et Stéfan Sinclair, *Hermeneutica. Computer-Assister Interprétations in the Humanities,* Cambridge, Massachusetts, MIT Press, 2016.

[6] *op.cit.* Sauf mention contraire, les informations sur les logiciels abordés dans l'état de l'art proviennent de cet article de Bénédicte Pincemin.

version récente. Nous pouvons noter un autre point commun, plutôt d'ordre pratique, à savoir son hébergement sur un serveur en ligne qui permet à l'utilisateur de ne pas avoir à télécharger un logiciel. Enfin, nous pouvons mentionner le logiciel **DtmVic** qui présente des fonctionnalités qui semblent similaires aux précédents logiciels. Au-delà des logiciels présentés dans cet article — et que nous ne pouvons tous abordés ici[7] — il en existe d'autres davantage axé sur la comparaison entre plusieurs version d'un texte comme **Hypermachiavel** ou encore sur l'étude d'un d'un corpus spécialisé à l'exemple de **BioTex** concentré sur les termes biomédicaux. Enfin, à l'aide du catalogue **Dataviz,** l'on peut voir que divers sites proposent une data visualisation sous forme de nuage de mots tout comme le fait Voyant Tools à la différence près que celui-ci est doté d'une quantité importante de méthode d'analyse du texte et de visualisation en découlant. Après avoir présenté le logiciel et avoir dressé un état de l'art sur la question de l'analyse de données textuelles, il s'agit à présent de justifier le choix de cet outil pour notre livrable, et ce au regard de notre projet de recherche.

II) Pourquoi Voyant Tools ?

Si notre choix s'est porté sur Voyant Tools, c'est d'une part parce qu'il correspond aux attentes que nous avons quant à notre mémoire qui porte sur le rôle des festivals dans le renouveau de la musique baroque aux XXe et XXIe siècles en France. Il s'agit de comprendre en quoi les festivals, véritables lieux de circulation culturelle, de création, de légitimation des artistes et du répertoire baroque, mais aussi lieux de diffusion, peuvent être considérés comme des passeurs culturels ? Et par quel moyens ont-ils participé à la redécouverte et l'installation de la musique baroque dans le paysage musical français des XXe et XXIe siècles ? Pour ce faire les archives (privées et publiques) sont mobilisées ainsi que des sources audiovisuelles. Mais c'est ici sur les sources de presse que nous nous concentrerons. En effet, selon le triptyque « production-médiation-réception » caractéristique de l'histoire culturelle, nous prêtons attention à la réception médiatique des festivals, à la manière dont la presse permet de produire puis de diffuser des représentations de ces derniers au grand public — et par extension de la musique baroque. Seulement pour étudier le discours médiatique, il faut être capable de saisir les discours tenus sur les festivals dans chacun des articles avant de pouvoir les confronter et ainsi monter en généralité dans notre analyse. C'est ce que nous avons tenté avec Voyant Tools. D'autre part, nous avons choisi cet outil pour la diversité de visualisations des données qu'il propose. Une diversité particulièrement intéressante puisqu'il nous semblait important de pouvoir synthétiser des analyses dans une visualisation afin de rendre ces données plus lisibles[8].
À cet égard, il convient de présenter les différentes possibilités de Voyant Tools qui ont retenu notre attention au regard des résultats que nous souhaitions obtenir.

---

[7] Dans cet article, Bénédicte Pincemin présente 7 logiciels de téxtométrie : DtmVic, Hyperbase, Hyperbase Web Edition, IRaMuTeQ, Lexico 5, Le Trameur et TXM. Nous ne pouvons pas tous les mentionnés et si mention il y a, elle s'appuie sur cet article dans la mesure où il n'est pas possible de tous les essayer et les prendre en main dans le cas présent.
Afin d'identifier la proposition de logiciel de textométrie, nous nous sommes également référée à une liste conçue par Huma-num : http://explorationdecorpus.corpusecrits.huma-num.fr/outils-logiciels-corpus-ecrits/.

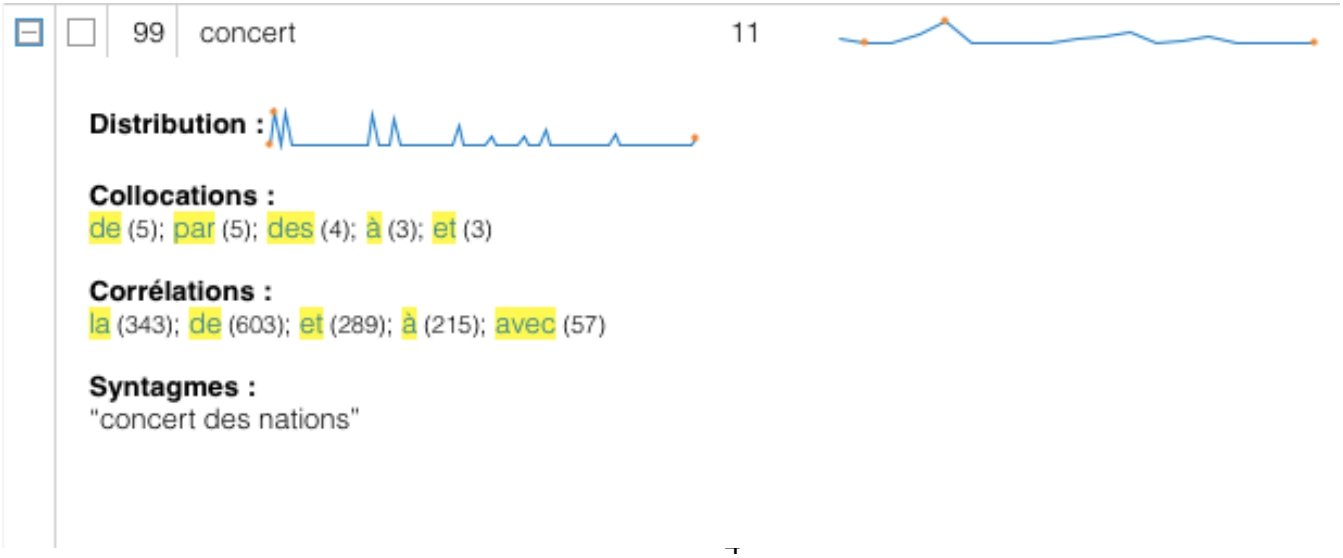
[8] « These representations transform complex data into relatively accessible forms; rather than being faced with raw numbers, percentages or lists, the user can explore the information visually. » - Tonkin Tourte, Gregory J. L. Harris, Tonkin Emma, and Harris Greg, *Working with Text : Tools, Techniques and Approaches for Text Mining,* ed. 2016, Chandos Information Professional Ser, Web, p. 53.

Les outils de visualisation

Tout d'abord, **MicroSearch** que nous avons retenu car, chaque document étant représenté par une colonne, il permet de repérer la récurrence d'utilisation d'un terme et la manière dont il est disséminé dans les textes.
Ensuite, l'outil **Tendances** avec lequel nous pouvons voir sous forme de graphiques (colonnes, aire, lignes, colonnes empilées) la fluctuation de l'utilisation d'un ou plusieurs termes.
Enfin, deux outils qui se présentent sous la forme de visualisation de mots. L'outil **Arbre de mots**, comme son nom l'indique, permet de visualiser à partir d'un terme ceux qui lui sont associés. De plus il est possible, en cliquant sur l'un d'entre eux, d'obtenir la suite de la phrase ou une arborescence d'autres mots qui lui sont associés. Et **Cirrus** qui, en s'appuyant sur la fréquence de représentation des mots dans le corpus ou un document unique, génère un nuage de mot. En cliquant sur un de ces mots, il apparaît dans les documents surlignés en jaune de manière à le replacer dans son contexte avec plus de facilité.

Les outils de tableaux

Justement, l'outil **Contextes** nous permet de voir les différentes utilisations qui sont faites d'un mot donné et en l'occurence choisi. Il se présente de la sorte : au milieu le terme recherché dans les documents. Ensuite, nous trouvons de part et d'autre de cette colonne centrale les morceaux de phrases qui l'entour nous permettant ainsi de replacer le mot dans son contexte. À une échelle plus globale, en cliquant sur l'un des rangs, le document concerné s'affiche avec le mot recherché surligné. Une fonctionnalité assez pratique que nous avons largement exploité comme nous le verrons dans la suite de ce livrable.
L'on doit une autre fonctionnalité intéressante, et complémentaire à celle présentée précédemment, à l'outil **Collocations**. Il s'agit d'un tableau qui couvre tout le corpus et met en lien un mot avec un autre qui lui est associé. Il est également possible d'effectuer cette recherche sur le mot-clé de notre choix.
Le dernier, et le plus général, est l'outil **Termes** qui montre l'intégralité des termes qui sont présents dans l'ensemble du corpus. Pour chacun d'entre eux, il est donné à voir les collocations et corrélations avec d'autres termes ainsi que le nombre pour chacune des relations et les expressions qui reviennent dans le corpus grâce à l'outil **Syntagmes**. Voici une capture d'écran avec l'exemple appliqué au mot « concert ».



Si les outils **Corrélations** ou **Collocations** sont également ouverts, dans le cas présent, en cliquant sur « concert », s'ouvre dans la fenêtre d'un de ces deux outils la recherche associée à ce terme.
Ajouté à cela, deux outils **Lecteur** et **Résumé**. Alors que le premier est une interface permettant d'avoir un accès au corpus, avec l'intégralité des textes ou un document en particulier. L'association de chacun d'eux à une couleur permet de se repérer dans le corpus et de visualiser l'emplacement des occurrences recherchées.
Ces fonctionnalités ont donc influencé notre choix. À présent, il convient de développer l'utilisation que nous avons fait de ce logiciel puis de mettre en perspective nos attentes avec les résultats obtenus.

III) Exploitations et résultats

Avant d'expliciter les différentes étapes de traitement et d'analyse de notre corpus, il convient d'en dire quelques mots d'introduction. Comme évoqué précédemment, nous avons choisi de traiter une partie de notre corpus de source de presse. Celui-ci est composé de 19 articles issus du journal *Le Monde*, soit une partie moindre de nos sources. Toutefois cette sélection semble rester suffisante pour expérimenter les fonctionnalités de Voyant Tools. Pour le replacer dans son contexte, tous les articles portent sur le Festival d'Ambronay. Un festival de musique baroque né en 1980 dans l'Ain et qui fait l'objet d'une étude de cas dans notre mémoire.

Dans une démarche à la fois de recherche de résultat mais aussi d'expérimentation et de découverte du logiciel, une fois les documents chargés dans Voyant Tools, la première chose à été de tester chacun des outils afin de cerner ceux qui permettent d'obtenir les résultats attendus — ceux présentés précédemment. C'est également au **Résumé** que nous avons prêté attention en premier lieu notamment pour voir quels étaient les mots principaux les plus représentés dans le corpus.

Mots les plus fréquents : musique (49); baroque (34); d'ambronay (30); festival (30); ambronay (24)

Ces mots font partie de la liste de mots-clés que nous avons établie afin de savoir quels termes recherchés lors de notre seconde étape qui n'est autre que l'utilisation de l'outil **Contextes** avec l'objectif d'étudier l'utilisation des mots suivants : musique, baroque, l'histoire, public, lieu et festival. C'est d'ailleurs ce dernier que nous prendrons pour exemple ici. De quelle manière le festival est-il représenté dans le discours médiatique et quelles idées lui sont associées ? À travers la recherche du terme « festival », nous avons pu voir dans quel contexte il est utilisé, ce que l'on dit du festival d'Ambronay et plus précisément ce qui retient l'attention des journalistes. Afin de pouvoir faire une analyse, nous avons créé un tableau indiquant le numéro du document, la citation de laquelle est issu le terme « festival » ainsi que des éléments d'analyse. Il suffit d'entrer le mot clé dans la barre de recherche pour voir apparaître les occurrences associées à leurs contextes.

Étant donné que le tableau est particulièrement important[9], en voici seulement un extrait ci-dessous.

[9] Nous avons relevé sur l'ensemble des mots que nous cherchions 89 occurrences et 27 en ce qui concerne les festivals.

Document	Contexte d'utilisation	Analyse
N°1	« a grandi sans bruit depuis dix ans un festival d'un intérêt exceptionnel . »	Adjectif qui qualifie le festival
	« En cette fin de semaine, le soleil dore le Revermont feuillu, et Pérouges accueille dans ses rues médiévales nombre des mélomanes qui font chaque jour la navette avec l'abbaye où se déroule le festival, sautant d'un Moyen Age à l'autre. »	Récit + associé avec le déroulement du festival
	« La musique ancienne reste bien la vocation première du festival, qui nous emmenait encore, samedi, aux offices de Florence, il y a tout juste quatre cents ans, pour le mariage de Ferdinand de Médicis et de Christine de Lorraine. »	Identité et programmation du festival
N°3	« Ainsi la reconstitution-restitution de ce San Ignacio d'Amazonie, écrit au début du XVIIIe siècle, exhumé au début des années 90 et donné dès 1996 à Santa Cruz de la Sierra (enregistré dans la foulée bolivienne par le label lorrain K. 617). Un événement majeur dans l'histoire de la musique. On comprend dès lors que le Festival d'Ambronay ait eu à coeur de le programmer par deux fois (1996 et 1999), que Ribeauvillé et Sarrebourg lui aient aussitôt emboîté le pas. »	Programmation du répertoire baroque latino-américain
	« Le festival continue jusqu'au 20 décembre. »	Information fonctionnelle (date)
N°4	« C'est bien le moins, dira-t-on, tant il est vrai que les compositeurs vivants n'ont jamais été autant en butte à la passion du public et des interprètes pour leurs prédécesseurs, valeurs sûres éprouvées par le temps.Mais le Festival d'Ambronay a aussi le souci de sortir des sentiers battus en proposant de véritables recréations dont le succès n'est jamais garanti. »	Identité artistique du festival + volonté de se distinguer + prise de risque

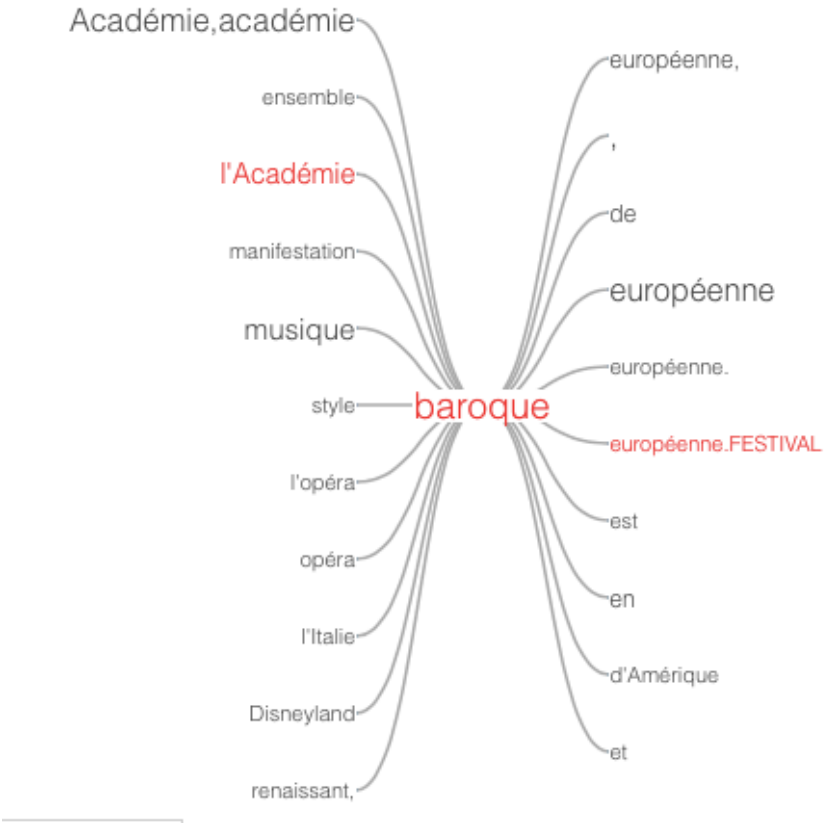
L'on peut voir dans la colonne « Analyse » que leur contenu n'est pas très intéressant en l'état, mais il le deviendra une fois que chaque utilisation sera mise en lien avec les autres. En cela nous pourrons monter en généralité, changer d'échelle d'analyse et identifier ce que met principalement en avant le discours médiatique à propos des festivals. Cet échantillon est assez représentatif dans la mesure où l'on retrouve à trois reprises la question de la programmation et de l'identité artistique du festival ce qui semble, même si l'analyse de l'entièreté de notre corpus pour le mémoire n'est pas encore aboutie, être un des points le plus mis en avant dans les articles. Ces articles sont aussi un des moyens utilisés pour assurer la publicité du festival. Ainsi l'on retrouve systématiquement, à quelques exceptions près, des informations que nous appelons « fonctionnelles ». Souvent situées à la fin des articles, elles donnent le prix des places, les lieux, les évènements phares de la programmation ainsi que les heures et dates des concerts comme il en est le cas dans l'exemple ci-dessus (document n°3). Ce sont aussi des éléments sur le déroulement du festival que l'on retrouve d'ailleurs davantage dans la presse nationale ou régionale par rapport à la presse spécialisée qui a tendance à plus mettre l'accent sur les performances des artistes en en faisant des critiques. Enfin, le document n°1 présente deux dimensions tout à fait intéressantes : les adjectifs attribués au festival qui permettent d'avoir une idée de sa réception, des avis qu'il cristallise ; et l'idée du récit. Certains des journalistes font parfois recours au récit de manière à réveiller l'imaginaire des lecteurs afin que leur apparaissent des images précises.

Cet outil est simple d'utilisation et les résultats sont assez satisfaisants. Le seul bémol est qu'il faut penser à toutes les variantes orthographiques des mots. Qu'il s'agisse du pluriel ou des article comme « d' », le logiciel les compte comme deux occurrences différentes. Cela implique donc soit un nettoyage des données en amont — mais cela ne viendrait pas fausser les résultats donnés par l'outil **Contextes** ? — soit de chercher les deux orthographes quitte à complexifié le résultat final.

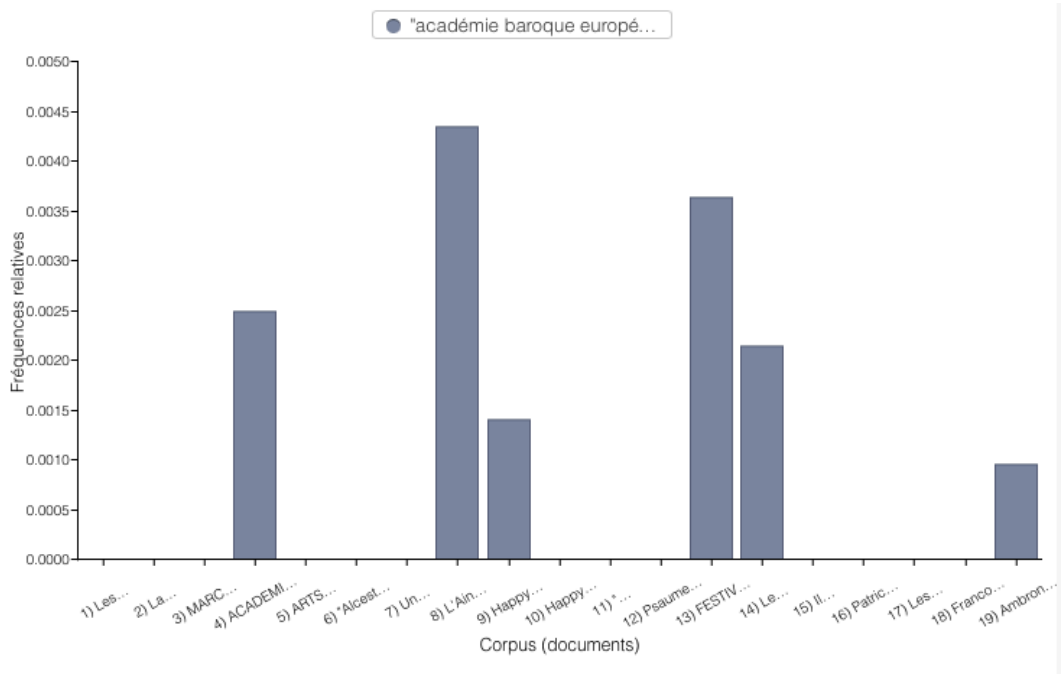
Les trois autres outils que nous avons utilisés nous ont davantage permis de faire des visualisations de nos données textuelles. En premier lieu, nous avons généré un nuage de mots avec l'outil **Cirrus**. Il est intéressant de voir que tous nos mots-clés ou presque figurent dans ce nuage de mots. Cependant, même s'il est possible de les ajoutés à la liste, de nombreux mots dits « vides » font partie de la visualisation mais ne sont pas pertinents pour l'analyse. Quand bien même nous les ajoutions dans cette liste des mots vides, pour avoir essayer, d'autres s'ajoutent et il est compliqué d'enlever ces mots dans leur intégralité. C'est là une limite de cet outil. Nous avons donc créé notre propre nuage de mots à partir de nos mots-clés afin de pouvoir les chercher plus rapidement dans les textes et voir lesquels sont les plus représentés. Afin de pouvoir comparer, voici les deux nuages de mots. Le premier généré automatiquement par Voyant Tools ; le second généré selon nos mots-clés.

Le second outil **arbre de mots** nous a permis d'identifier les mots principaux qui sont liés dans l'exemple ci-dessous au mot « baroque ». Les résultats ne semblent pas toujours pertinents et exploitables. Par exemple, certains mots sont associés à des articles et ce même si nous plaçons ces articles dans la liste des «mots vides ». Pourtant, dans le cas présent les résultats sont représentatifs

de l'impression que nous avons eu à la lecture des articles même si il y a des doublons et des articles ou une virgule qui n'apporte rien à l'analyse.



Le dernier outil que nous avons utilisé est **Tendances** car pour compléter les visualisations précédentes, il est intéressant de pouvoir mesurer la fluctuation de la représentation d'un thème à travers le temps. Ici nous avons fait un essai avec l'Académie baroque européenne qui nous a semblé à la lecture faire l'objet de beaucoup d'articles. En effet, sur 19 articles, 6 d'entre eux portent entièrement ou en partie sur cette académie.



Enfin, nous ne pouvons évoquer les quelques outils qui se sont finalement avérés peu utiles au regard des résultats qu'ils ont donné. **MicroSearch** a certes permis de voir la la dissémination des termes dans les documents mais il n'est pas possible de les retrouver dans les textes concernés. De ce fait, il n'a pas fait l'objet d'une analyse particulière. De même pour les outils **Corrélations** et **Collocations** puisque, au regard de l'objectif recherché, ils donnaient des résultats beaucoup moins exploitables que l'outil **Contextes** permettant de replacer le mot dans son contexte d'utilisation par exemple.

Conclusion

Pour conclure, le logiciel Voyant Tools présente de nombreux avantages à commencer par l'aspect pratique puisqu'il n'est pas nécessaire de télécharger le logiciel qui est accessible en ligne. De plus, son interface est intuitive et pratique — par exemple il est possible de changer la dimension des fenêtres et tous les outils sont disponibles dans chaque fenêtres, nous pouvons donc travailler à notre guise. Hébergé par Huma-num pour la version française, le contenu importé dans Voyant Tools est protégé ce qui est particulièrement important car les chercheurs peuvent être amenés à entrer des données qu'il n'est pas forcément possible de diffuser en l'état. Le logiciel présente une grande variété d'outils d'analyse de corpus et de visualisation et permet ainsi à l'utilisateur de tout faire à partir du logiciel sans avoir besoin d'entrer ses données dans plusieurs logiciels.

En ce qui concerne les limites du logiciel, les données nécessitent un travail de nettoyage en amont pour que certains résultats soient exploitables mais il nous semble que certains des outils ne fonctionneraient pas correctement si les documents sont nettoyés, nous pensons en particulier à l'outil **Contextes**. Aussi même si Voyant Tools facilite tout de même le travail d'analyse textuelle de l'utilisateur, une étude « humaine » des textes est nécessaire afin de ne pas laisser de côté certains éléments. Prenons l'exemple des adjectifs ou expressions qui se rapportent au festival d'ambronay. Dans les résultats, n'apparaîtront que les termes reliés au mot « festival » mais deux cas en seront exclus. D'une part, les termes qui caractériserons le festival sans que le mot

« festival » soit présent dans la phrase. D'autre part, les résultats ne prendront pas en compte les synonymes ou expressions remplaçant « festival » à l'exemple de « manifestation », « événement » ou encore un surnom du festival.

Toutefois, nous envisageons de passer le reste de nos sources dans le logiciel afin de faciliter et systématiser le repérage de nos mots-clés ainsi que leur utilisation dans les différents textes. C'est donc l'outil **Contextes** qui sera particulièrement employé. Associé à une lecture et analyse par nous-même, les résultats devraient être assez exhaustifs.

Bibliographie

Damon Mayaffre, Bénédicte Pincemin, Céline Poudat, « Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse des discours », Langue française, Armand Colin, 2019, Les outils informatiques au service des linguistes, pp.101-115. hal-02419199

Pincemin, Bénédicte, « Sept logiciels de textométrie », 2018. halshs-01843695.

Tonkin Tourte, Gregory J. L. Harris, Tonkin Emma, and Harris Greg, *Working with Text : Tools, Techniques and Approaches for Text Mining,* ed. 2016, Chandos Information Professional Ser, Web

Sitographie

« Exploration de corpus : outils et pratiques », http://explorationdecorpus.corpusecrits.huma-num.fr/outils-logiciels-corpus-ecrits/.

Huma-Num. La TGIR des humanités numériques, https://www.huma-num.fr/services-et-outils/traiter#texte.

« Voyant Tools help », http://voyant.tools.huma-num.fr/docs/#!/guide/about;

« Introduction à Voyant Tools », https://github.com/aurelberra/voyant_tools/blob/master/tutorial/voyant_tools_intro_fr.md