Lily Grumbach

MHN 1

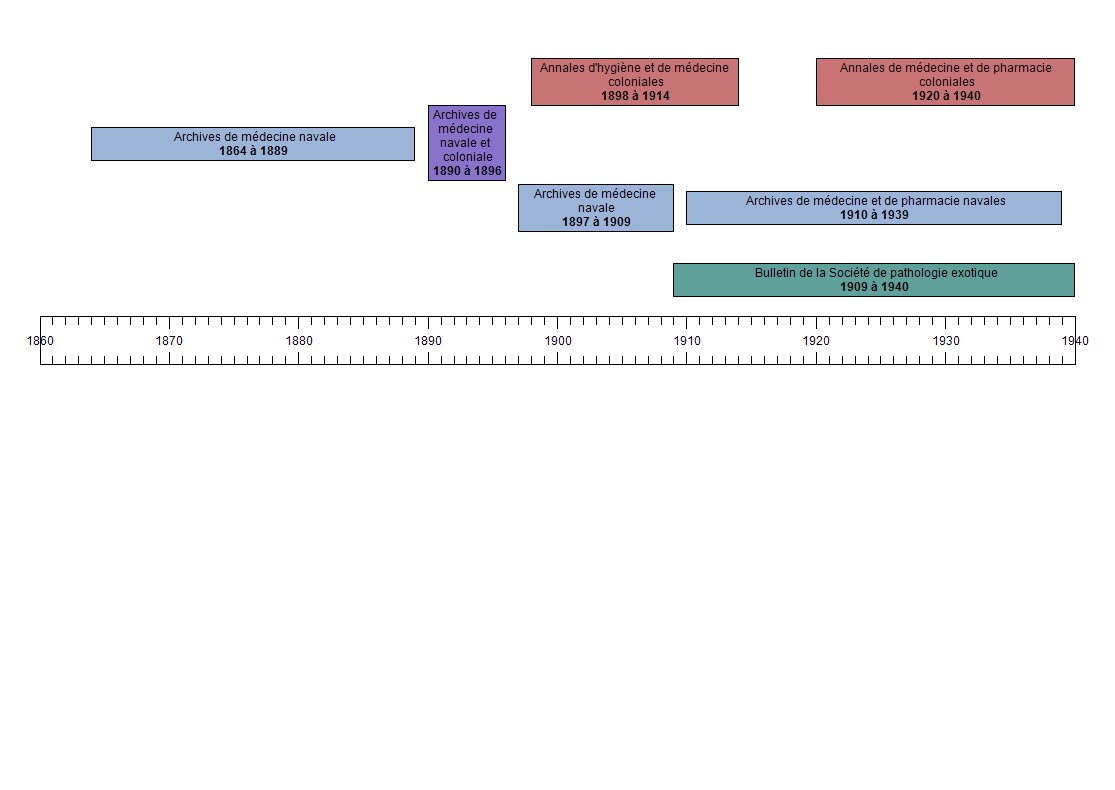
**Traitement automatique de la langue**

Le développement de nouveaux paradigmes scientifiques (théorie des germes, rôle des parasites), l’expansion et la stabilisation des domaines coloniaux et l’institutionnalisation des « sciences coloniales »[[1]](#footnote-1) marquent les années s’écoulant de 1880 à 1914. A travers le monde se déploient alors laboratoires, sociétés savantes et écoles de médecine dédiées à l’étude des affections appelées « exotiques », « tropicales », « coloniales » en fonction de sa position dans le champ politique. Dans ce cadre, la communication scientifique entre métropole et colonies mais aussi entre les colonies et avec les autres empires coloniaux prend une place de premier ordre.

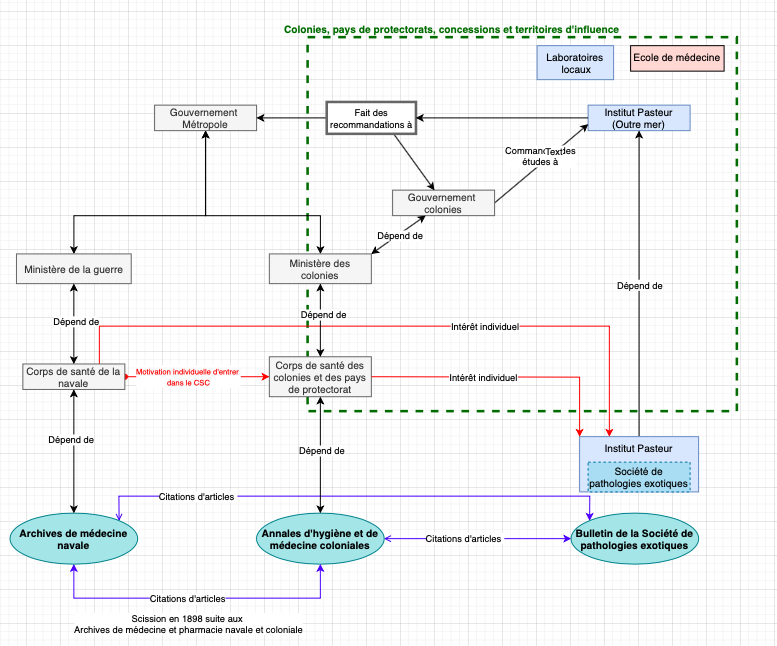
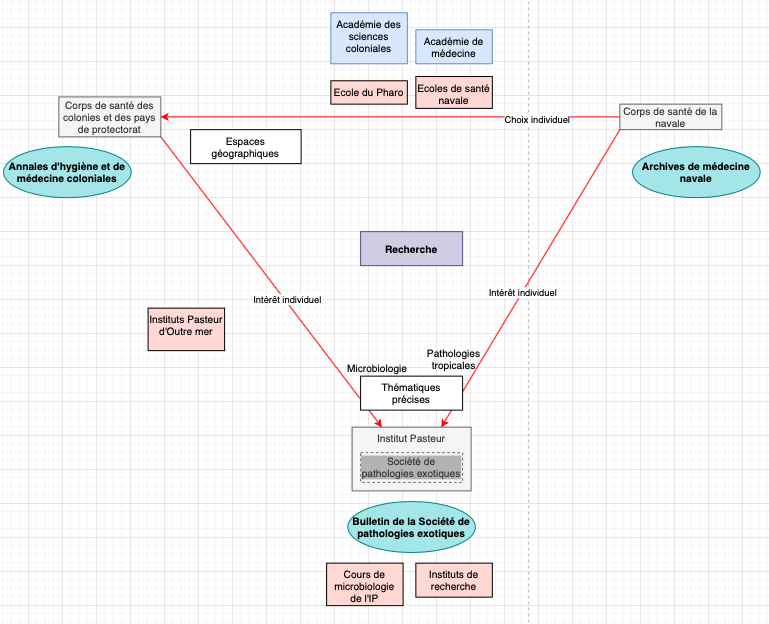
Pour notre mémoire, nous étudions trois revues liées par leurs acteurs et leurs thématiques : les Annales d’hygiène et de médecine coloniale (AHMC), les Archives de médecine navale (AMN) et le Bulletin de la Société de pathologie exotique (BSPE). Si l’étude des acteurs fait l’objet d’un autre rendu, nous avons souhaité, grâce aux méthodes de traitement automatique de la langue vues en cours, étudier l’évolution du vocabulaire employé afin d’en estimer les thématiques. En effet, travaillant cette année uniquement avec les titres des articles nous n’avons pas souhaité mettre en place un topic modeling qui n’aurait pas eu assez de matière pour constituer un modèle pertinent.

L’idée principale était alors de saisir l’évolution des thématiques abordées au cours du premier vingtième siècle (1898-1940) et, si le temps imparti nous le permettait, étudier depuis 18989, c’est-à-dire depuis la création du Corps de santé des troupes coloniales. En effet, l’histoire des revues institutionnelles que nous étudions dans ce devoir est le résultat d’une scission administrative et politique entre le corps de santé de la marine et celui des colonies au fil des années 1890. La question que nous nous posons est alors de saisir, à partir des articles originaux de cette revue (nous avons au préalable retiré l’ensemble des sections de la revue relevant de la bibliographie, du bulletin officiel, des listes de promotion) comment se distinguent les thématiques et lieux traités par les deux revues.

Par souci pratique et pour pouvoir maîtriser nos revues, nous avons mené notre étude sur les 10 premières années des AHMC et des AMN, c’est à dire de 1898 à 1908. Cette période est en effet clé dans notre étude. Tout d’abord, elle marque les premières années de « spécialisation » respective à l’issue de la scission de 1898. Ensuite, elle précède la création du Bulletin de la Société de pathologie exotique. Cette dernière constitue un majeur car cette société, créée sous l’égide d’Alphonse Laveran et d’Emile Roux à l’Institut Pasteur non seulement capte les principaux médecins du corps des troupes de santé coloniales mais en plus se positionne très rapidement comme fer de lance des recherches adressées aux enjeux sanitaires dans les colonies et sur les mers.



Généalogie des revues étudiées.



Nous présenterons donc ici les deux stratégies que nous avons souhaité chaque fois mettre en place et quels en étaient l’intérêt pour notre mémoire.

1. Reconnaissance d’entités nommées

La première tentative de rendu TAL a été d’appliquer à notre corpus une reconnaissance d’entités nommées. Cette démarche consistait à entraîner, sur une partie de notre base de données constituée de l’ensemble des titres de nos revues, une NLP à partir d’annotations réalisées sur Doccano. L’objectif était non seulement de désambiguiser les entités de nos données mais aussi de proposer de nouveaux labels d’annotation correspondant à notre corpus afin de l’enrichir pour notre rendu de l’an prochain. Un autre objectif, à plus court terme était de corriger les entités nommées relatives à un lieu afin de les passer par la suite sur QGIS pour un rendu de TAL. Cette seconde partie du rendu fait l’objet d’un autre devoir dans le cadre du séminaire TAIS dirigé par Eric Mermet.

Afin de répondre à ces objectifs nous avons donc réalisé une annotation sur Doccano en ayant recours à des labels préexistants du NLP ainsi que

Ainsi, nous avons eu recours aux labels suivants :

|  |  |  |  |
| --- | --- | --- | --- |
| **Label** | **Signification** | **Intérêt** | **Difficultés rencontrées** |
| PATH | Pathologie | Étudiant des revues médicales, il nous a semblé primordial de donner voix aux noms de pathologies qui peuvent par ailleurs souvent porter le NLP de base à confusion en raison de noms de maladie faisant parfois référence à un lieu. Par exemple : « Bouton d’Alep » | De nombreuses fois, les maladies étaient référées d’autres formes grammaticales ne relevant donc pas d’une entité nommée à strictement parler.  Par exemple : ‘paludéen’ plutôt que ‘paludisme’. |
| ZOO-ENTO | Zoologie-entomologie | Depuis la découverte de Laveran du rôle transmetteur des anophèles, une attention croissante a été donnée aux moustiques, animaux et protozoaires. | Comme pour CHEM-LABO, les données que nous avons utilisé |
| SHS | Sciences humaines et sociales | Nous avons considéré dans la catégorie « SHS » les noms se rapportant ou découlant d’une discipline des sciences humaines et sociales, c’est-à-dire apportant un savoir autre que strictement médical ou biologique. La géographie et l’ethnologie en sont les principaux apports. | Nous avons dans ce cadre annoté des noms d’ethnies mais restions assez perplexes dans des cas de classification d’individus sur la base de caractères stéréotypés.  Par exemple : « les noirs ». La question était alors de savoir dans quelle mesure cette désignation généraliste découlait d’une considération ethnologique, d’une seule observation ou bien d’un stéréotype. |
| CHEM-LABO | Chimie et laboratoire | Ce label a pour intérêt d’estimer le passage d’un régime médical de la clinique à celui de la biomédecine.  Étudier les entités nommées de chimie nous aurait aussi permis de faire une étude des pratiques de laboratoire. | Trop peu présent sur la période que nous avons annoté, il serait pertinent d’annoter le Bulletin de la Société de pathologie exotique ou bien les mêmes revues après leur rattachement aux services de pharmacie pour avoir plus d’exemples. |
| LOC | Localisation | Le principal intérêt de la ré-annotations sur Doccano des LOC était de désambiguïser certains noms ambivalents, notamment de maladies (cf PATH) ainsi que d’affirmer la différence avec les GPE. |  |
| GPE | Geopolitcal Entity | Entités géopolitiques.  De nombreuses entités GPE étaient annotées comme LOC ou bien mal tokenisées. C’était notamment le cas pour les noms composés indochinois |  |
| ORG | Organisation |  | La grande majorité des noms d’organisation étaient liés à des lieux. Étant donné l’objectif premier du rendu de cartographier les contributions aux articles, nous avons fait le choix de prioriser les GPE et LOC aux ORG pour nous éviter une étape supplémentaire dans la géolocalisation.  Néanmoins, il serait préférable à l’avenir de prioriser les ORG pour être le plus proche du texte. |

Pour cela, nous avons mis en place un 4-fold cross validation afin de nous assurer d’avoir un NLP robuste. Néanmoins, il nous semble à postériori que le rapport entre la quantité de données entrainées (1898-1908) et l’ensemble de notre corpus (1898-1940) nous menait à surentraîner notre NLP sur des des données trop restreintes et finalement assez peu représentatives de l’ensemble de notre corpus. Néanmoins, il nous semble qu’il serait pertinent d’appliquer cette NLP aux 10 années précédentes (1889-1897) correspondant aux années durant lesquelles le corps de santé de la marine et celui des troupes coloniales écrivaient à la même adresse, à savoir aux Archives de médecine navale et coloniale. Cela pourra être développé pour le mémoire de l’an prochain que nous comptons préparer dès cet été.

Nous avons mis en place une manière de calculer la précision et l’accuracy de nos annotations par entité. Ayant eu des problèmes avec notre boucle for, nous n’avons pu la déployer pour nos prédictions. Nous avons donc fait le calcul manuellement à partir de deux extraits aléatoires de 15 titres d’articles.

Une image contenant texte, reçu, capture d’écran

Description générée automatiquement

Nous pensons toutefois avoir dans notre démarche quelque peu confondu l’apport d’une REN par rapport à du topic modelling. La précision de la REN ne pouvait en effet pas toujours correspondre thématiques que nous aurions souhaité encoder.

1. **Word Embeddings**

L’objectif dans cette partie était de vectoriser les mots de chaque revue pour en saisir les différences. Nous avons opté pour cette méthode en raison de son accessibilité et du fait que nous l’avions pratiquée en cours, notamment avec M. Plancq. Pour cette deuxième partie, nous avons souhaité vectoriser les mots de nos revues en comparant d’une part les 10 premières années des AHMC et AMN entre elles mais aussi aux années de la revue des Archives de médecine navale et coloniale. Cette démarche nous aurait permis de saisir les positions des mots en fonction des revues et de saisir lesquels étaient les plus couramment utilisés ensemble.

Le premier jupyter notebook, en tant qu’exploration des textes par matrices de mots, fait dans ce cadre-là office d’une première exploration du texte qu nous avions poursuivis en Word2Vec.

Conclusion

Si mes tentatives n’ont pas pu aboutir en raison de maladresses techniques à des heures de travail peu recommandables, j’ai tout de même pu acquérir une meilleure compréhension des enjeux et fonctionnements de deux majeurs domaines du traitement automatique de la langue que sont la reconnaissance d’entités nommées et les word embeddings. Mon objectif pour le mémoire final de l’an prochain serait de réaliser pleinement ce que j’ai initié pour ce devoir (en m’assurant d’avoir des backups … évidemment ‼).

1. Pour des études plus précises sur l’histoire de ces «spécialités », voir Laurence Monnais-Rousselot, *Medecine et colonisation: l’adventure indochinoise 1860-1939*, Paris, CNRS Editions, 1999, 489 p ; Helen Tilley, *Africa as a living laboratory: empire, development, and the problem of scientific knowledge, 1870-1950*, Chicago, University of Chicago Press, 2011, 496 p ; Emmanuelle Sibeud, *Une science impériale pour l’Afrique? la construction des savoirs africanistes en France, 1878-1930*, Paris, Ecole des hautes études en sciences sociales, 2002, 356 p. [↑](#footnote-ref-1)