Anatomie, animaux, vocabulaire de la vivisection

Construire des ressources lexicales pour visualiser une thématique dans un corpus littéraire

Philippe Gambette, Tita Kyriacopoulou,
Nadège Lechevrel, Claude Martineau
UPEM, laboratoire LIGM & ANR/DFG BIOLOGRAPHES, FMSH

Cet article propose une méthodologie d'annotation et de visualisation, en vue de l'analyse, de textes d'un corpus littéraire sur la thématique de l'expérimentation animale. Elle se fonde notamment sur l'extraction du vocabulaire relatif à cette thématique qui concerne plus précisément l'anatomie, les animaux ainsi que l'expérimentation. Pour cela, nous combinons deux outils, Unitex et TreeCloud, afin, d'une part, d'enrichir des ressources linguistiques pour la langue française présentes dans la distribution d'Unitex, et d'autre part de visualiser les thématiques d'intérêt au sein du corpus, au fil du texte, ou de manière synthétique.

Introduction

Le projet de recherche Animalhumanité visait à réunir chercheuses et chercheurs en littérature, sciences du vivant et informatique pour des travaux sur l'expérimentation et la fiction mettant l'animalité au cœur du vivant, en valorisant à la fois les fonds des collections du musée Fragonard et le fonds ancien de la bibliothèque de l'EnvA, École nationale vétérinaire d'Alfort. En raison de la non-disponibilité en version numérique des ouvrages du fonds ancien, nous nous sommes concentrés sur les descriptions des pièces du musée Fragonard référencées sur le site de la Bibliothèque interuniversitaire de santé de Paris[1], ainsi que sur un corpus constitué d'ouvrages, déjà numérisés et

1 http://www.biusante.parisdescartes.fr/histoire/images/index.php?mod=a&orig=enva.

disponibles en mode texte, suggérés par les collègues littéraires impliqués dans le projet. Ce corpus a été traité par deux outils afin d'y mettre en valeur les thématiques d'intérêt du projet : Unitex, un analyseur de corpus fondé sur des ressources linguistiques et TreeCloud un logiciel issu de la textométrie qui produit la visualisation d'un texte sous la forme de nuage arboré[2]

Présentation du corpus

Plusieurs références ont été transmises par les chercheuses et chercheurs en littérature impliqués dans le projet de recherche *AnimalHumanité*, en vue de constituer un corpus textuel sur la thématique de l'expérimentation animale, ou plus précisément de la vivisection. Une première phase du travail a consisté à rechercher des sources numérisées, disponibles au format texte, pour ces références. Des sources variées ont été utilisées : *Wikisource*, le site du *Labex OBVIL*, *Gallica*, *Frantext*, *The Montaigne Project*, les *Bibliothèques virtuelles humanistes*, le *Centre Flaubert*, *Les classiques des sciences sociales*, le *Musée de La Fontaine* et *archive.org*. Finalement, un corpus de 34 textes a été constitué sur le principe de l'« échantillon de convenance »[3], c'est-à-dire en combinant les besoins thématiques avec les contraintes de disponibilité.

Ce corpus, dont la majorité des textes date du XIX^e^ siècle, est disponible sur la page http://eclavit.univ-mlv.fr/animalhumanite. Il s'agit d'un corpus de taille réduite (près de 3 Mo et 500 000 occurrences), relativement hétérogène, en particulier du point de vue de la longueur des textes (certains n'étant que des extraits) ou de la langue utilisée (romans, ouvrages scientifiques, œuvres en français du XVI^e^ siècle non modernisé).

2 Philippe Gambette. *User Manual for TreeCloud,* 2010. http://www.treecloud.org/DOWNLOADS/ManualTreecloud.pdf ; Philippe Gambette, Jean Véronis, "Visualising a Text with a Tree Cloud", IFCS'09 (Proceedings of the International Federation of Classification Societies 2009 Conference), *Studies in Classification, Data Analysis, and Knowledge Organization* n°40, 2010, p. 561-570. https://hal-lirmm.ccsd.cnrs.fr/lirmm-00373643/fr/.

3 Mark Algee-Hewitt, Mark McGurl, "Between Canon and Corpus: Six Perspectives on Twentieth-Century Novels", *Stanford Literary Lab Pamphlet* n° 8, 2015, http://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.

Enrichissement des ressources pour l'annotation des textes par Unitex

Une annotation automatique basée sur des ressources lexicales et des motifs grammaticaux

Unitex est un logiciel libre multilingue et multiplateforme[4] d'analyse de corpus qui fait appel à des ressources linguistiques (dictionnaires et grammaires locales). Il permet en particulier de localiser des *motifs*, c'est-à-dire des mots ou groupes de mots qui correspondent à un patron combinant des contraintes lexicales ou morphosyntaxiques. Ces contraintes peuvent s'exprimer sous forme d'un automate, comme celui illustré en figure 1[5]. Les motifs détectés dans le texte sont alors *annotés*, c'est-à-dire que des balises sont automatiquement ajoutées pour indiquer leur appartenance à une catégorie donnée.

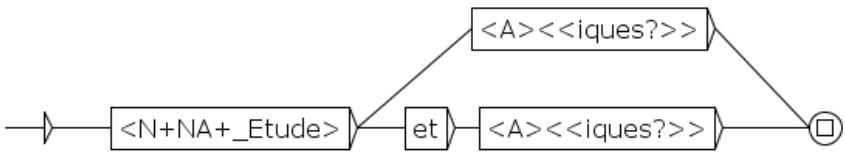

Figure 1 : Automate construit avec Unitex pour reconnaître des expressions du type nom commun, ou groupe nominal, lié à la thématique « études », suivi d'un adjectif se terminant par « iques » ou par « et » suivi d'un adjectif se terminant par « iques ».

À partir de son interface, ce logiciel permet de ne traiter qu'un seul texte à la fois. Pour traiter notre corpus, nous avons donc développé un script en Perl qui appelle directement le cœur du logiciel Unitex et qui permet de produire pour chacun des textes, un texte balisé avec des annotations. Ce texte annoté est ensuite traité par le programme Perl qui génère une page web dans laquelle chaque annotation est surlignée d'une couleur qui indique sa catégorie. En plus du texte annoté, l'outil fournit aussi l'ensemble des motifs

4 Unitex dispose d'un site internet : http://unitexgramlab.org.

5 Sébastien Paumier, De la reconnaissance de formes linguistiques à l'analyse syntaxique, thèse de doctorat, Université de Marne-la-Vallée, 2003, https://hal.archives-ouvertes.fr/tel-01687029.

reconnus, classés par fréquence décroissante ou par catégorie.

Nous avons donc annoté le corpus à l'aide de ce script et obtenu la page web disponible à l'adresse http://eclavit.univ-mlv.fr/animalhumanite à partir des 23 catégories indiquées en figure 2.

Catégorie	Nb d'occ.	Motifs diff.
Outil_Chirurgical	305	25
Médical	497	71
Anomalie	63	8
Biologie	1631	285
Chimie	1294	244
Profession	701	30
Expérimentation	1075	72
Homme_Animal	1089	58
Étude	4675	793
Forme_Verbale	877	296
Personne	1318	391

Catégorie	Nb d'occ.	Motifs diff.
Animal_domestique	533	74
Mammifère	497	123
Oiseau	361	91
Insecte	732	110
Reptile	98	22
Animal	136	37
Pré_Animal	85	17
Cat_Animal	251	74
Partie_Corps	2432	304
Partie_Corps_Animal	367	43
Fluide_Corporel	384	33
Institution	37	3

Figure 2 : Synthèse des 23 catégories recherchées dans le corpus, avec pour chacune le nombre de motifs différents détectés (3204 au total) et d'occurrences reconnues (19438 au total).

Un enrichissement des ressources lexicales utilisées

Nous disposions déjà dans nos dictionnaires de traits de type animal, parties du corps, etc. Cependant, certaines ressources étaient insuffisantes ou inadéquates, ce qui nous a amenés à les compléter à partir de notre corpus, à partir de règles linguistiques ou à partir de la base de données des pièces du musée Fragonard.

En ce qui concerne l'ajout de traits plus précis que ceux déjà présents dans nos ressources, nous avons créé des catégories « animal domestique », « mammifère », « oiseau », « insecte », « reptile » afin d'augmenter la finesse de nos repérages. Un trait « parties du corps animal » a également été ajouté au trait « parties du corps », pour repérer des mots comme « pattes » ou « bec ».

Nous avons également complété certains dictionnaires par des entrées supplémentaires. Un certain nombre de ces nouvelles entrées, sur les parties du corps animal, les animaux ainsi que des anomalies médicales par exemple, proviennent d'une analyse arborée, montrée en figure 3, des titres des pièces du musée Fragonard présents dans une base de données. L'arbre rapproche les mots qui apparaissent fréquemment dans les mêmes titres, et il est coloré en fonction des diverses catégories thématiques proposées (animaux en bleu, parties du corps en rose, anomalies anatomiques en rouge, entités nommées en vert).

La méthode de recherche de motifs caractéristiques de certains traits, implémentée par Unitex, est aussi utilisée pour ajouter de nouvelles entrées à nos ressources lexicales. Par exemple, nous avons constaté lors de la lecture du corpus que plusieurs parties relatives aux descriptions de débats scientifiques font apparaître des mots composés avec des noms suivis d'adjectifs se terminant par « ique », tels que « propriétés mécaniques » ou « sciences biologiques ». Nous avons donc construit l'automate illustré en figure 1 pour repérer ces expressions. À partir de la liste des résultats obtenus, nous avons ajouté à nos dictionnaires ceux qui pouvaient être rattachés à la description d'études scientifiques, en les associant à un trait « étude », et en ajoutant les codes employés par Unitex pour reconnaître automatiquement les formes fléchies, au pluriel, par exemple. Pour de plus amples explications sur ce graphe on pourra consulter les sections 4.3 et 4.7 du manuel d'Unitex[6].

Finalement, plus de 2500 entrées spécifiques ont ainsi été ajoutées à nos dictionnaires.

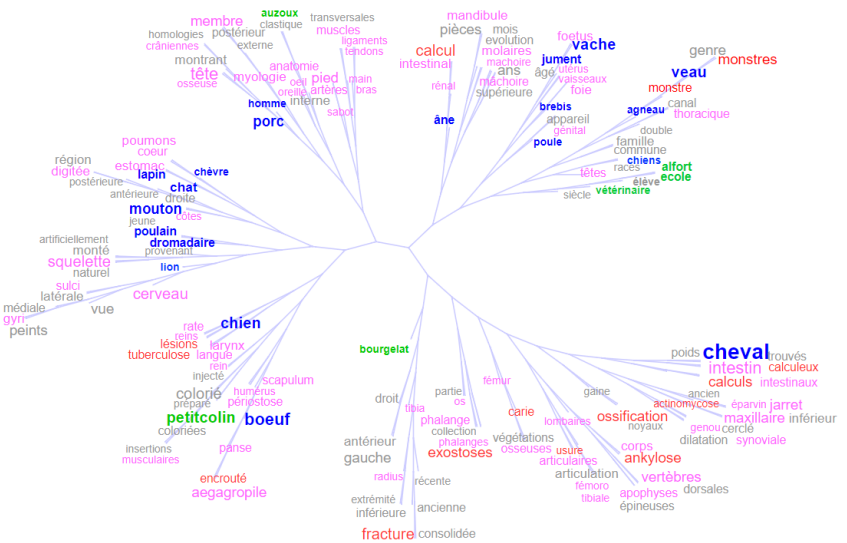

Figure 3 : Nuage arboré des mots présents dans mots dans au moins 15 descriptions de pièces du musée Fragonard de l'EnvA, parmi les 3084 recensées dans la collection EnvA BIU Santé. Disponible sur http://treecloud.univ-mlv.fr/treecloud-linker/fragonard.html en version interactive avec des liens vers le formulaire de recherche dans la collection EnvA.

6 Sébastien Paumier, Claude Martineau, Unitex 3.1, *Manuel d'utilisation*, 2016 http://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-fr.pdf.

Visualisations et analyses du corpus

Notre corpus est assez large et contient plusieurs œuvres où seuls quelques extraits concernent les thématiques du projet *AnimalHumanité*. Nous avons donc commencé par extraire d'un nuage arboré construit sur l'ensemble du corpus la liste des termes les plus fréquents liés à la thématique de l'expérimentation animale : « expérience », « expériences », « supplice », « anatomie », « sang », « sanglants », « émotion », « pitié », « horreur », « peur », « scalpel », « aiguilles », « éther », « phosphore », « poisons », « morte », « mort », « horrible », « barbare », « assassin », « injecter », « enfermer », « souffrir », « mourir ». Nous avons alors observé le voisinage de ces mots à l'aide de l'extraction de concordances (10 mots avant et 10 mots après) et de leur visualisation en nuage arboré.

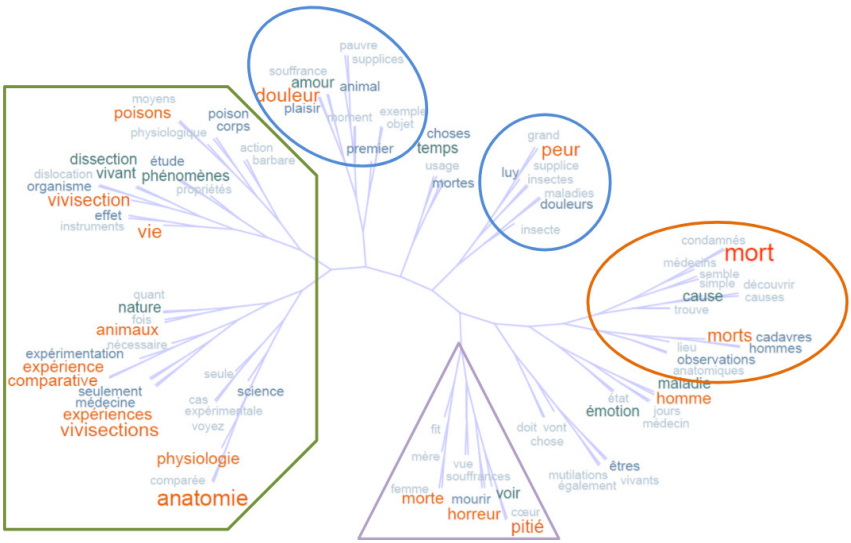

Figure 4 : Nuage arboré des 100 mots les plus fréquents (hors mots vides) dans le voisinage des termes de la catégorie « expérimentations ».

Dans ce nuage arboré montré en figure 4, le sous-arbre de gauche est consacré aux démarches de recherche en sciences du vivant. Il mêle par exemple les disciplines (anatomie, physiologie, médecine), procédés d'étude (vivisection(s), dissection, expérience(s), expérimentation, dislocation), les objets d'étude (organisme, vie, vivant, nature, animaux) et certains moyens utilisés (poison(s), instruments).

Deux sous-arbres en haut de la figure 4 sont consacrés au ressenti de l'animal (douleur(s), peur, supplice(s), souffrance), alors qu'en bas un seul

sous-arbre mêle un lexique lié au féminin (mère, femme, morte) et à des émotions ressenties en observant (vue, voir) les expérimentations (horreur, pitié, cœur). Enfin, un sous-arbre à droite associe à la thématique de la mort (mort(s), cadavres) un vocabulaire essentiellement dénué d'émotion (le mot apparaît plutôt à côté de maladie et homme dans un autre sous-arbre) et plutôt lié à des analyses scientifiques de cadavres (cause(s), observations anatomiques).

Il est aussi possible de se concentrer sur des catégories de mots issues des ressources lexicales, en construisant par exemple avec Unitex les concordances des termes issus de la liste des parties du corps animal. En les visualisant de nouveau à l'aide d'un nuage arboré montré en figure 5, le nombre important d'occurrences de « patte », au pluriel et au singulier, apparaît. D'une façon générale, il y a beaucoup de « pattes » dans *Les Scènes de la vie privée* de Balzac, *Les Sabots de Noël* d'Haraucourt, *L'ennemi des lois* de Barrès et *L'insecte* de Michelet. Dans les textes, les pattes sont surtout celles des chiens, des chats et des chevaux, et le thème des pattes ficelées revient souvent car il est emblématique de la privation de liberté et de la souffrance : être ligoté et retenu pour l'expérimentation.

Les mots « bras » et « pattes » apparaissent à proximité dans l'arbre : un retour au texte montre qu'il peut y avoir des bras et des pattes pour diverses raisons. Ici, dans le texte de Michelet, l'homme et l'animal sont confondus dans une métaphore filée de « *l'insecte géant qu'on appelle cerf-volant, l'un des plus gros de nos climats, masse noire et luisante aux cornes armées de superbes pinces* » où il est tantôt prisonnier, tantôt Roméo[7]. Michelet s'émeut également devant les longs bras d'enfants d'un puceron[8].

L'anthropomorphisme facilite aussi la vulgarisation : Michelet évoque ainsi les dents et la bouche pour décrire les mandibules d'une fourmi[9]. Enfin, la proximité inattendue dans ce corpus des mots « yeux » et « cœur » dans

7 « Il la palpait de ses pattes et de ses bras tremblotants. Il parvint à la retourner, tâtonna (très probablement il ne voyait plus), pour bien s'assurer si elle vivait. Il ne pouvait s'en séparer ; l'on eût juré qu'il avait entrepris, lui mourant, de ressusciter cette morte. » (Jules Michelet. *L'Insecte*, Paris, Librairie Hachette, 1858, http://corpus.biolographes.eu/titre.php?id=185).

8 « Jeté sur le dos, il étalait un ventre très-gros, une très-petite tête informe qui ne semble qu'un suçoir, et remuait toutes ses pattes qu'on eût dit plutôt de longs bras d'enfants. Au total, un être innocent, et qui n'inspire aucune répugnance. » (Jules Michelet, *ibid.*)

9 « Je profitai avec hâte de l'attitude pénible où je tenais ma fourmi : je regardai son visage. Ce qui désoriente le plus et lui donne un aspect étrange, ce sont principalement les dents ou mandibules, placées en dehors de la bouche, et partant l'une de droite, l'autre de gauche, horizontalement, pour se rencontrer ; les nôtres sont verticales. Ces dents en avant menacent et semblent présenter le combat. Cependant, comme nous l'avons dit, elles ont des usages pacifiques et servent aussi de mains. Derrière ces dents apparaissent de petits filets ou palpes, à l'entrée de la bouche. Ce sont en réalité comme de petites mains de la bouche, qui palpent, manient, retournent ce qu'on y apporte. Du front partent les antennes, autres mains, mais du dehors, mobiles à l'excès, sensibles, des mains électriques. » (Jules Michelet, *ibid.*)

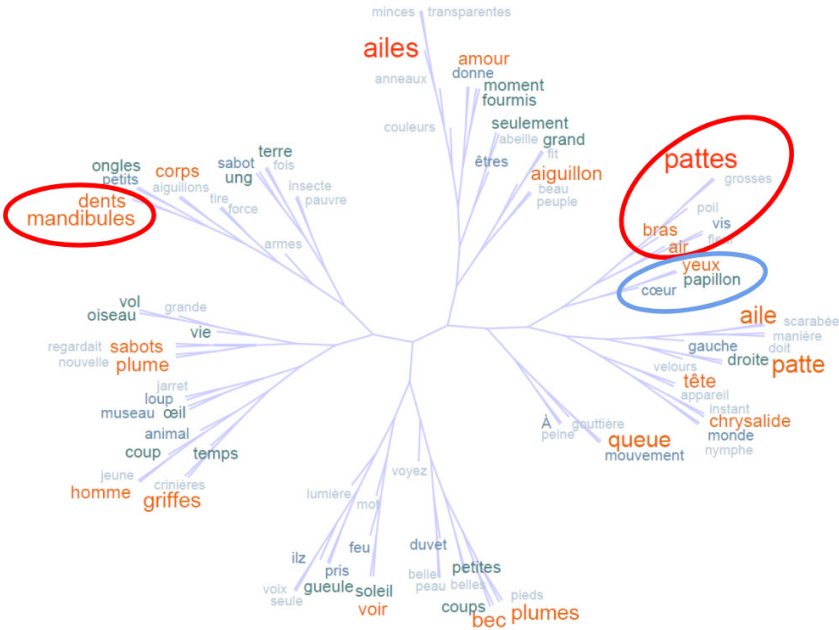

Figure 5 : Nuage arboré des 100 mots les plus fréquents dans les contextes (10 mots avant, 10 mots après) des mots de la catégorie « parties du corps animal ».

la figure 5 s'explique en partie par un extrait où Michelet s'interroge sur l'humanité des insectes[10].

Ainsi, les parties des corps humain et animal sont interchangeables à souhait dans les textes littéraires : c'est le pouvoir effroyable et monstrueux de la vivisection qui fait de l'homme une bête, et révèle l'humanité de l'animal.

Conclusion

La méthode présentée ci-dessus a permis de définir 23 catégories correspondant aux intérêts des personnes impliquées dans ce projet à l'interface de la littérature, des sciences de la vie et de l'informatique. Notons toutefois que les textes antérieurs au XIX^e^ ne sont que très faiblement représentés dans le corpus en raison de leur moindre disponibilité et des difficultés à traiter le français de l'époque avec les ressources linguistiques et les outils informatiques dont nous disposons.

En appliquant des méthodes d'analyse linguistique et statistique à des corpus littéraires, nous facilitons la fouille du corpus en ligne pour les collègues, en fonction de leurs intérêts. Cela permet de mettre en relation des extraits de texte avec des pièces du musée par exemple dans le cadre d'une application mobile. Par ailleurs, les ressources lexicales construites peuvent être utilisées sur d'autres applications ou corpus.

10 « Point de regard dans ses yeux. Nul mouvement sur son masque muet. Sous sa cuirasse de guerre, il demeure impénétrable. Son cœur (car il en a un) bat-il à la manière du mien ? Ses sens sont infiniment subtils, mais sont-ils semblables à mes sens ? Il semble même qu'il en ait à part, d'inconnus, encore sans nom. » (Jules Michelet, *ibid.*)