Les variations des sens

Le texte, le hors texte et le corpus

Geoffrey Williams

Professeur en sciences du langage,
Université de Bretagne-Sud, UMR Litt&Arts

Résumé : Ce texte pose la question de la nature des humanités numériques, ainsi que des disciplines et des outils qu'elles impliquent. Si la désignation « Humanités numériques » est récente, certains des constituants, notamment la linguistique de corpus et la TEI, Text Encoding Initiative, ont des racines dans les années quatre-vingt, alors que d'autres outils comme les CAQDAS, Computer Assisted Qualitative Data Analysis Software, sont plus récents. Après un brève historique de ces trois derniers, le texte examinera le rôle de ces disciplines et outils à travers deux exemples : la numérisation des pièces de théâtre de Louis de Boissy et du Dictionnaire universel d'Antoine Furetière dans sa version élargie de 1701.

Mots-clés : CAQDAS ; XML-TEI ; Dictionnaires historiques ; Théâtre ; linguistique de corpus ; BaseX

1 Introduction

L'interaction entre dictionnaire, corpus et texte est complexe, et essentielle. Un dictionnaire repose forcément sur une décontextualisation de la langue. En partant des écrits les plus anciens, Clarisse Herrenschmidt a très bien décrit en 2007 la décontextualisation par l'écrit, un écrit qui nous permet de stocker et transmettre des données linguistiques, mais en perdant tout contact avec le contexte de création. Dans un monde textuel, nous avons tendance à oublier la nature même de l'écrit et ses désavantages, du fait de notre fascination pour ses avantages évidents. L'origine des dictionnaires se trouve dans cette décontextualisation, motivée par la nécessité de donner des clés d'accès aux données, que ce soit par un dictionnaire de codification des caractères comme en Chine, ou par la dénomination des dieux comme en Grèce. Cependant, tant que l'oralité est restée le moyen principal d'échanger des informations, le rôle du dictionnaire a été limité.

Le passage des gloses sur un texte médiéval aux glossaires puis, à partir du XVI[e] siècle, aux dictionnaires est l'histoire d'une décontextualisation croissante. Pendant le Moyen-Âge, le monde des monastères, gardiens de l'écrit, était non seulement un monde textuel, mais un monde post-latin où la pratique croissante des langues vernaculaires rendait l'accès au texte plus difficile. La solution se trouvait dans des gloses, les annotations des mots importants dans les pages de l'ouvrage, des explications fixées dans le contexte d'emploi du texte. Cependant, cette contextualisation avait ses inconvénients du fait même que chaque explication était confinée au texte d'origine. Les glossaires ont fourni une solution, en gardant un lien avec le contexte textuel par leur nature thématique. Les *catholicons*, collections de glossaires, ont rompu ce lien thématique. Les dictionnaires ont entériné cette rupture par leur nature onomasiologique, qui remplace toute organisation conceptuelle par l'emploi simple de l'alphabet. Certes, à partir du XVII[e] siècle, en commençant avec les travaux d'Estienne et le dictionnaire monumental de Furetière, un contexte élargi a refait son apparition, avec des exemples et des citations. Cependant, cette fenêtre s'est vite fermée avec le prescriptivisme, qui utilise les citations pour donner du prestige à certains mots plutôt que pour les expliquer. Les exemples, quand ils étaient encore là, étaient forgés et non représentatifs de la réalité des emplois. Le dictionnaire a été mis sur un piédestal comme outil de référence, au lieu d'être un outil reflétant l'emploi réel. Nous oublions trop facilement qu'un dictionnaire ne reflète pas l'emploi actuel des mots, mais uniquement la façon dont les mots étaient utilisés à une époque donnée, et ceci après une médiation fondée sur un processus d'introspection sur le sens des mots.

La question qui se pose est donc de retrouver le contexte. La réponse se trouve dans la linguistique de corpus, *corpus linguistics*. La linguistique de corpus est une discipline qui est née en Grande Bretagne dans les années soixante, bien avant ce que nous appelons maintenant les humanités numériques. Cependant, les outils développés par la discipline sont au cœur des humanités numériques, que sont également venus nourrir le balisage des textes selon les recommandations de la Text Encoding Initative (TEI) et plus récemment le développement des Computer Assisted Qualitative Data Analysis Software (CAQDAS).

Dans le texte qui suit sera posée la question de la nature des humanités numériques, ainsi que des disciplines et des outils qu'elles impliquent. Pour ce faire, on s'appuiera sur des exemples de projets qui lient disciplines humanistes et outils numériques.

2 Que peuvent être les humanités numériques ?

Les humanités numériques (dorénavant HN) sont une discipline (ou un regroupement de disciplines et d'approches méthodologiques), assez récente dans le paysage de la recherche française. On les pratique parfois sans le savoir, comme Monsieur Jourdain faisait de la prose. Il en est ainsi de la linguistique de corpus et de la TEI (Test Encoding Initiative) dans la construction et la gestion des corpus, entreprises menées pendant des années dans une sorte de désert, et avec un rattachement disciplinaire assez questionnable à la (toutefois) grande chapelle de la section 7 du CNU, et à propos desquelles on s'est rendu compte récemment du rôle pionnier qu'elles ont joué dans l'émergence de ce qu'on appelle aujourd'hui les HN. Avec cette prise de

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

conscience, elles sont devenues plus acceptables dans le monde par ailleurs clos des études littéraires et historiques, où elles ont souffert, au moins jusqu'à présent, d'une marginalisation certaine.

Pour comprendre de quelle façon les HN sont perçues, je me suis livré à un petit exercice de sondage auprès de quelques collègues bretons, auxquels j'ai demandé de se positionner par rapport à la discipline ; ce sondage (peu formel) a été mené dans le cadre d'un effort de structuration, mené par la Maison des Sciences de l'Homme de Bretagne, et qui cherche à faire émerger un réseau des HN dans la région. Assez significativement, les réponses sont venues des collègues en sciences, tandis que le silence a été quasiment total du côté des humanités, même si je sais que certains de nos historiens se sont déjà engagés dans la voie des HN.

Les exemples suivants, puisés parmi les réponses reçues, montrent la grande confusion qui règne au sujet de ce que peuvent être les HN :

— *Je ne travaille pas directement en « humanités numériques », mais il y a des intersections entre ce domaine et l'interaction homme-machine (IHM).*
— *Je n'ai pas de collaboration particulière avec le service informatique, mais nous utilisons actuellement la visioconférence (salle immersive) pour nos enseignements sur plusieurs sites.*
— *Est-ce que cela a un rapport avec les TICE ?*

Cerner les perceptions et les pratiques s'est donc révélé peu révélateur au regard de la question posée.

Quand on veut en savoir un peu plus sur un sujet, un point de départ, le point de départ pour la plupart de nos étudiants, est de regarder dans Wikipédia. On y lit que « Les humanités numériques [...] sont un domaine de recherche, d'enseignement et d'ingénierie au croisement de l'informatique et des arts, lettres, sciences humaines et sciences sociales [1] ».

Ce qui est donc évident est qu'il y a l'informatique d'un côté et l'ensemble des disciplines qu'on appelle les sciences humaines et sociales de l'autre. La notion d'« humanités » est prise de manière très large, ce qui peut être problématique quand on sait que, tout en étant regroupées sous la même dénomination, les sciences sociales et les humanités ont des modalités de fonctionnement très différentes et bénéficient de traitements institutionnels et politiques fort divergents (ainsi, le financement de la recherche va très largement vers les sciences sociales, et très peu vers les humanités classiques).

La version anglaise de Wikipedia est légèrement différente :

> The Digital Humanities are an area of research, teaching, and creation concerned with the intersection of computing and the disciplines of the humanities.
>
> The humanities include ancient and modern languages, literature, philosophy, religion, and visual and performing arts such as music and theatre.

1. https://fr.wikipedia.org/wiki/Humanites_numeriques [consulté le 15 juin 2020].

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

> The humanities that are also sometimes regarded as social sciences include history, anthropology, area studies, communication studies, cultural studies, law and linguistics [2].

La définition anglaise restreint donc beaucoup plus le champ des HN, ce qui est probablement plus proche de la réalité puisque les sciences sociales « dures », comme les sciences politiques et la sociologie, ont leurs propres paradigmes, davantage empiriques que la textualité qualitative des humanités. Ce qui est probablement le plus important, c'est l'idée de collaboration de l'informatique en tant que discipline d'ingénierie, qui travaille avec les chercheurs en humanités et en fonction de leurs besoins – afin de faire avancer leurs paradigmes. Une collaboration ne devrait pas être simplement une source de financement supplémentaire pour l'ingénierie informatique, en oubliant les besoins et les pratiques des humanistes. Malheureusement, les financeurs européens et nationaux sont davantage fascinés par la technique que par la sagesse.

Ce que révèlent ces différentes définitions, et dans une certaine mesure les réponses au sondage mentionné plus haut, c'est le problème du rapport entre les deux champs que les HN mettent en contact. Que se passe-t-il précisément ? Risque-t-on une prise de pouvoir de l'informatique sur les humanités ? Ou bien assiste-t-on à un apport d'outils essentiels pour faciliter la recherche ? Ces outils sont-ils liés uniquement à l'ingénierie, ne concernant que des activités particulières comme l'édition numérique ou la gestion des archives, ou bien bouleversent-ils profondément les paradigmes ?

Le spectre d'une prise de pouvoir de l'informatique sur les humanités peut facilement se concrétiser si les humanistes ne font pas l'effort de s'approprier la démarche. Dans beaucoup d'institutions, on assiste à l'arrivée de techniciens qui connaissent bien l'informatique, mais qui ne communiquent pas, ou peu, avec les communautés qu'ils sont censés servir. Il est vrai que l'ouverture d'esprit doit avoir lieu des deux côtés, et que dans beaucoup de cas les collègues humanistes ont tendance à traiter les informaticiens/ techniciens comme de simples prestataires de services.

Pour ce qui est de l'apport, par l'informatique, d'outils essentiels pour faciliter la recherche, on est dans le domaine de l'évidence. La question est plutôt celle d'un effort insuffisant d'adaptation de ces outils, en grande partie développés ailleurs, pour la recherche en lettres, langues ou histoire, ainsi que celle de leur appropriation. Deux excellents exemples sont les concordanciers employés dans la linguistique de corpus et les outils d'aide à la recherche qualitative, les CAQDAS. Ces derniers ont été créés par et pour les sociologues, mais sont parfaitement adaptés pour la recherche en lettres et langues. Cependant, peu de collègues de ces disciplines savent qu'ils existent, et qu'il suffit de les adopter pour des bénéfices évidents.

Il n'est, enfin, pas moins vrai que ces outils sont particulièrement utiles pour certaines activités ou domaines disciplinaires. En tant que responsable d'un master en métiers du livre et des humanités numériques, je sais que les HN apportent énormément de choses à un monde de l'édition en mutation. Mais, à ne regarder les HN que sous l'aspect de leurs applications professionnelles, on risque de créer des diplômés

2. https://en.wikipedia.org/wiki/Digital_humanities [consulté le 15 juin 2020].

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

technocrates, qui ne sauront pas véritablement comprendre le monde des livres, ni suivre l'évolution des rapports entre le lecteur et les données.

Une fois de plus, la question s'avère donc celle du point d'équilibre entre informatique et humanités. Pour essayer de cerner celui-ci, je prendrai comme exemple cette discipline de la linguistique de corpus, que j'ai déjà mentionnée, et la pratique de la TEI qui est venue nourrir mes travaux de lexicographe. Comment s'ouvre-t-on à de nouvelles pratiques, comment choisit-on des outils, sans perdre son âme ?

3 De la TEI au *Corpus Linguistics*

La destruction des œuvres écrites par des incultes n'est pas une nouveauté. Le savoir fait peur et les obscurantistes ont toujours existé. Cependant, l'ennemi numéro un reste le temps et la dégradation physique des supports. Dans les deux cas, la conservation numérique peut garantir l'accès aux contenus en cas de perte et donne la possibilité d'accéder aux documents sans les manipuler.

La conscience des avantages d'un stockage numérique est apparue dès les débuts du SGML (Standard Generalized Markup Language). Toutefois, à l'époque les solutions étaient prescriptives. Le principal apport de la TEI a été, dès lors, de fournir des solutions descriptives, capables de s'adapter à tout type de texte, de toutes formes et en toutes langues.

La première réunion TEI a eu lieu à Vassar College, Poughkeepsie, NewYork State en 1987. À partir de là, des deux côtés de l'Atlantique s'est mise en place une collaboration entre chercheurs en sciences humaines et informaticiens, qui a permis l'évolution des recommandations (*guidelines*), le développement d'outils de visualisation – logiciels, systèmes, feuilles de styles… – ainsi que la création d'un site riche en aides et conseils. En 1999, la décision de transformer la TEI en consortium a été prise, afin d'en assurer la pérennité. Depuis, la TEI s'est imposée comme le modèle par excellence pour le traitement des manuscrits et des textes étudiés par les humanités. La TEI est essentielle pour la numérisation des archives, des textes littéraires et historiques, des corpus de langues. Elle a pu passer du SGML à l'XML et évoluer en même temps que les évolutions technologiques, tout en étant un moteur de ces changements. Depuis 2012, quelques membres du consortium ont fondé les Digital Humanities Awards afin de valoriser les travaux entrepris. Il est donc légitime de dire que les humanités numériques sont nées avec la TEI.

Le développement de la TEI a été surtout lié à une question de stockage, de mise à disposition et de visualisation de textes historiques et littéraires. En langues et linguistique, un autre chemin a été emprunté, celui de l'analyse des corpus avec des concordanciers afin d'extraire des informations linguistiques. La gestion d'un corpus (voir plus loin pour une définition du mot « corpus » dans le sens de la linguistique de corpus) est très différente de la gestion des textes, ou collections de textes. Les corpus sont le plus souvent créés à partir de sources déjà numériques. Ils sont de très grande taille – plusieurs millions de mots. Il s'agit de textes de sources différentes et de genres différents, d'où la nécessité de taxonomies et de partitions. Néanmoins, avec une quantité croissante de textes historiques librement disponibles, nous devrions pouvoir

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

créer de véritables corpus de textes anciens afin de regarder l'environnement de la création d'une œuvre littéraire et pas simplement l'œuvre elle-même. La linguistique de corpus pourrait donc bientôt bénéficier aux lettres et langues.

4 *Corpus Linguistics* / linguistique de corpus

La linguistique de corpus, *corpus linguistics*, est une discipline qui est apparue en Grande-Bretagne dans les années soixante, mais qui a vraiment pris son essor dans les années 1980 avec la démocratisation de l'informatique par le biais des ordinateurs de bureau. L'origine de l'intérêt des chercheurs en langues et linguistique pour l'informatique est plus ancienne.

Pour le moment, nous allons utiliser le terme « linguistique de corpus » comme équivalent au terme anglais *corpus linguistics*, mais la réalité est le plus souvent autre. Il y plusieurs tendances en linguistique de corpus, certaines plus littéraires, d'autres plus marquées Traitement Automatique de la Langue. Le cœur de la discipline reste l'analyse de la langue en tant qu'outil de communication humaine. Ses textes fondateurs sont ceux de John Rupert Firth (1957), à partir desquels s'est développée ultérieurement l'école contextualiste ou école néo-firthienne.

C'est surtout pendant la guerre de 1939-1945 que le gouvernement britannique a compris l'utilité de faire travailler ensemble des mathématiciens spécialisés dans la cryptographie et des linguistes. Les travaux entamés à Bletchley Park ont donc ouvert la voie à une étude linguistique informatisée. Le résultat fut le premier grand ordinateur, Colosse, créé pour casser les codes allemands, ce qui suscita évidemment un intérêt pour l'informatique du côté des linguistes qui collaborèrent à ce travail.

Les collaborateurs de Bletchley Park n'ont pas pu raconter leur histoire, puisque de la guerre chaude nous sommes passés immédiatement à la guerre froide. Cependant, suite à ces travaux et grâce au génie d'Alan Turing, l'informatique moderne était née.

D'abord essentiellement dépendants des mathématiciens, certains linguistes ont pu accéder aux ordinateurs. Tel fut le cas à Edimbourg, ou le linguiste Angus McIntosh, un ancien de Bletchley Park, a recruté le jeune John Sinclair. C'est donc à Edimbourg que Sinclair a créé le premier corpus d'étude linguistique dans le cadre du projet OSTI (voir Krishnamurthy, 2004), projet terminé à Birmingham après sa nomination en tant que professeur. Bien qu'oublié dans un tiroir pendant plus de trente ans, le rapport OSTI signale le début de la linguistique de corpus en fixant la plupart des termes employés de nos jours. À la fin du projet, Sinclair et son équipe ont un peu délaissé les corpus afin de travailler sur l'analyse de discours avant de reprendre avec le projet COBUILD (Collins Birmingham University International Language Database) dans les années 1980 (Sinclair, 1987) .

Le succès du projet COBUILD a suscité un très grand intérêt et la création d'un grand corpus très largement diffusé [3], le British National Corpus, BNC. Le BNC est le premier grand corpus entièrement balisé en parties de discours et lemmatisé. C'est aussi le premier corpus construit selon les normes de la TEI. Avec la popularité

3. Le corpus COBUILD, devenu par la suite le Bank of English, est d'un accès limité et cher, sauf pour les membres de l'université de Birmingham.

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

du BNC, l'arrivée massive d'ordinateurs personnels sur le marché et l'apparition de concordanciers évolués comme WordSmith Tools [4], la linguistique de corpus a été très largement adoptée par les chercheurs en langues, surtout pour des travaux portant sur l'enseignement des langues ou sur la lexicographie.

Ce qu'on appelle la linguistique de corpus en France est une discipline assez différente de la *corpus linguistics* puisqu'elle est issue de tendances beaucoup plus littéraires. Une partie de la méthodologie peut être employée, mais l'esprit n'est pas nécessairement le même et il s'agit dans beaucoup de cas de simplement travailler sur corpus (Williams, 2006). Ceci ne rend pas l'approche moins intéressante, mais il est utile d'éviter les confusions. Beaucoup de chercheurs travaillent sur corpus sans nécessairement faire de la linguistique de corpus (Williams, 2010).

Pendant les années 1960, les possibilités offertes par l'informatique ont été explorées par d'autres linguistes, notamment par Paul Imbs et Bernard Quémada en France et Nelson Francis et Henry Kučera aux États Unis. Tous ont eu une grande influence sur la linguistique de corpus, sans faire de la linguistique de corpus. Les premiers répondent à l'idée d'Imbs de créer un grand dictionnaire historique du français, le Trésor de la langue française. Ce projet entamé par Imbs était novateur à plus d'un titre, mais encore plus remarquable était le recours à une base textuelle informatisée comme source de citations, Frantext. À l'origine, la base Frantext était limitée aux textes du XVIII^{e} et du XIX^{e} siècle, mais, de nos jours, il s'agit d'une très grande collection de documents numérisés, certes de tendance littéraire, qui traverse les siècles et l'histoire de la langue française. Cependant, Frantext n'est pas un corpus, c'est une collection de textes. Il est possible de construire des ensembles, des « corpus de travail » selon la terminologie des littéraires, mais il ne s'agit pas de corpus dans le sens de *corpus linguistics*. Par ailleurs, peu de textes sont disponibles librement et l'accès à la base en ligne est limité. Seule l'interface Frantext est disponible pour mener des requêtes, ce qui limite les questions susceptibles d'être posées à celles qu'ont imaginées ses concepteurs. Or, un véritable travail en *corpus linguistics* suppose la possibilité d'utiliser ses propres outils afin de trouver des réponses aux questions spécifiques de chaque chercheur.

5 Le *corpus linguistics*

Comme nous avons déjà vu, le *corpus linguistics* est à l'origine surtout un phénomène britannique, qui s'est développé en adoptant les perspectives ouvertes par le linguiste Firth et en utilisant des ordinateurs afin d'étudier le contexte large. La base de la philosophie de Firth réside dans la formule « You shall know a word by the company it keeps » (Firth, 1957, p. 11), autrement dit par le contexte et la collocation.

Dans la pensée de Firth, la collocation, en tant que phénomène textuel, est omniprésente. Les deux notions clés derrière la linguistique de corpus sont le contexte de culture, hérité de Malinowski, et le contexte de situation. Les deux signifient qu'il n'y a pas de sens hors contexte. Construire un corpus signifie mettre le contexte de création et d'emploi des textes au centre des préoccupations du chercheur. Comme

4. http://www.lexically.net/wordsmith/ [consulté le 13 juin 2020].

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

l'a dit plus tard Sinclair, il faut faire confiance aux textes (Sinclair, 2004). La base de la linguistique de corpus est le corpus, et le corpus est composé de textes. Il ne faut jamais perdre de vue les textes et les choix opérés dans leur sélection. Ainsi un corpus est « a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research » (Sinclair, 2005).

Pour comprendre cette définition, il faut faire l'effort de comprendre le raisonnement sur lequel elle repose, ce qui signifie lire le texte en entier. Malheureusement, peu le font. Il ne suffit pas de lire les textes fondateurs de la discipline, il faut essayer de les comprendre sans imposer son propre modèle, ce qui a été trop souvent le cas du TAL et de la linguistique de corpus française. Dans une telle définition, il est important de comprendre la notion de texte authentique, qui refuse les données inventées et prises hors contexte. Un texte est forcément une unité textuelle créée dans un but de communication et dans un contexte repérable. Un corpus est basé sur des choix (Engwall, 1994). Ce ne doit pas être le fruit du hasard, issu d'une collecte de textes sur le Web assistée par ordinateur.

Les idées de Firth ont donnée naissance à deux grandes tendances, la grammaire systémique et fonctionnelle de Halliday, et la *corpus linguistics* de Sinclair. Ces deux tendances peuvent être largement complémentaires, mais le point de départ de toute l'analyse de Sinclair est le texte, et le mot en contexte (Sinclair, 1991). Le *corpus linguistics* exige de construire des corpus selon des principes de représentativité et selon des choix linguistiques explicites (Sinclair, 2005). La base de l'analyse, c'est la concordance. À la différence d'autres disciplines, l'informatique est un outil d'analyse, mais pas l'objet de l'étude. L'idée est de pouvoir revenir constamment au texte. Une analyse en parties de discours peut être utile, mais en réalité une telle approche impose déjà un système et une analyse, au lieu de chercher le système dans le corpus. En *corpus linguistics*, l'analyse est forcément inductive, *corpus-driven* (Tognini-Bonelli, 2001), et l'introspection n'a absolument pas sa place, ce qui écarte également les tendances cognitives puisque, comme l'analyse critique de discours (qui est tout sauf critique), celles-ci imposent leur idées sur le corpus au lieu d'interpréter les données du corpus. Ainsi, le *corpus linguistics* est une discipline qui possède à la fois des méthodologies et une éthique de travail. Il ne suffit pas simplement de prendre un concordancier et de se lancer.

L'école qui est issue des collaborations autour de John Sinclair, souvent appelée l'école de Birmingham, a généré une riche littérature sur les questions lexicographiques (Sinclair, 1987) et les analyses lexicographiques sur corpus (Hanks, 2013), la prosodie sémantique (Louw, 1993), l'amorçage lexical (Hoey, 2005), le discours (Teubert, 2010) et d'autres études importantes autour du mot en contexte.

La leçon de base de Firth était l'importance du contexte, mais nous travaillons en permanence avec des contextes réduits ou inexistants. Un corpus est un assemblage de textes, donc le texte est pris en masse hors contexte. Si le corpus a été assemblé correctement, et si les métadonnées ont été correctement renseignées, le contexte initial peut être retrouvé. Cependant, l'avantage d'un corpus bien construit est d'avoir accès au contexte de la langue, à l'usage à un moment donné. En effet, comme l'a

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

souligné Tognini Bonelli (2001), un corpus de référence, de taille suffisante, peut être vu comme représentatif de la langue dans le sens de Saussure. Si l'occurrence est la parole, un instant potentiellement imparfait d'emploi de la langue, la masse des données qui peut être visualisée avec un concordancier constitue la langue et donne la possibilité de tirer des conclusions générales sur l'emploi de la langue à un moment donné.

6 Les CAQDAS

Comme il a été dit plus haut, un linguiste de corpus ne doit pas oublier qu'un corpus est constitué de textes et doit rester toujours proche des textes. Cependant, tous les supports ne sont pas analysables par les concordanciers. Avant d'attaquer la construction et l'analyse d'un corpus, il faut pouvoir faire un travail d'analyse qualitative sur d'autres textes. Pour ce faire, les CAQDAS représentent un moyen de réaliser des études qualitatives rigoureuses sur des grandes quantités de données en forme de texte, mais aussi en image, et les PDF sont des images.

Les CAQDAS sont apparus dans les années 1980 (Brugidou et Le Roux, 2005 ; Leavy, 2014 [5]). Ils ont été créés surtout par et pour les sociologues, soucieux de traiter de grandes quantités de données, en particulier des entretiens, de manière rigoureuse. Un bon CAQDAS est capable de gérer de très nombreux formats, des textes et des images, et aussi du son et de l'audiovisuel. Il y a quelques outils gratuits, mais le marché est dominé par des logiciels commerciaux très puissants comme N-Vivo et Atlas.Ti. Pour mes travaux, j'utilise Atlas.Ti pour Mac. Plutôt que de décrire les fonctionnalités hors contexte, je détaillerai les avantages de cet outil dans des études sur le contexte dans les textes.

7 Au fond du XVIII^{e} siècle : un cas d'étude

Le groupe LicORN – Linguistique de corpus et des ressources numériques – est un groupe pluridisciplinaire ayant participé à de nombreux projets ANR (Agence nationale de la recherche) et PCRD (Programme cadre pour la recherche et le développement) sur la création et l'exploitation des corpus à des fins différentes. Au cœur de ses préoccupations reste la linguistique de corpus, ou plutôt le *corpus linguistics*, et la lexicographie.

Les projets autour de la citoyenneté (InTune PCRD 6), des écrits scientifiques (Scientext, ANR) ou de l'extraction multilingue de termes (METRICC, ANR), pour ne prendre que trois exemples, ne posent pas trop de difficultés en termes de modélisation du corpus. Le corpus correspondant à du langage contemporain, les difficultés résident dans les critères de sélection, dans l'obtention, la gestion des contenus, et, évidement, dans l'analyse. Ces difficultés deviennent en revanche beaucoup plus importantes quand on essaie de travailler sur des textes du passé, comme c'est le cas avec les projets Boissy et Furetière.

5. Voir aussi le site dédié de l'Université de Surrey : http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/about/index.htm [consulté le 10 juin 2020].

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

Ces deux projets sont issus de la rencontre entre deux chercheurs de disciplines différentes : moi-même, directeur de LiCoRN, linguiste de corpus et lexicographe, et une chercheuse sur le théâtre français du XVIII^{e} siècle, Ioana Galleron. Le point de départ était d'analyser le mot « sacré » dans le contexte du XVIII^{e} siècle français et l'influence de l'anglais du XVII^{e} siècle, plus précisément de John Milton (Galleron, 2015). L'idée de ces travaux, toujours en cours, était de comparer les textes, les définitions dans les dictionnaires des deux époques et un corpus de textes en anglais et en français des deux périodes. Ces travaux exploratoires ont montré les difficultés dans la construction d'un corpus équilibré, vu la quantité importante de textes disponibles pour l'anglais et la pauvreté des textes librement accessibles pour le français. Pour l'anglais, de nombreux textes sous format XML sont disponibles sur le site Oxford Text Archive ; pour le français, les barrières sont multiples. Ces barrières illustrent les difficultés auxquelles est confronté le développement des HN en France, et la grande nécessité de faire sauter les verrous qui nous empêchent d'accéder librement à des données patrimoniales.

Ce n'est pas que les textes manquent pour le français. En France, nous avons la chance d'avoir une vaste quantité de textes librement disponible en formats PDF sur le site Gallica de la BnF. Cette offre est inégalée dans le monde anglo-saxon, où les textes clés, comme le dictionnaire de Samuel Johnson, ne sont pas facilement accessibles en PDF. Cependant, pour créer un corpus, il faut des textes en XML, et c'est là où les difficultés commencent.

La base Frantext est intéressante, mais la quantité de textes librement disponibles *via* le site de la CNRTL (Centre national de ressources textuelles et lexicales) est limitée. Les autres ne sont accessibles que *via* l'interface de Frantext qui est peu *user-friendly* (conviviale), et qui de plus n'est accessible que depuis un ordinateur d'une université, ce qui suppose que le chercheur est universitaire et dans son bureau. Les requêtes sont inévitablement limitées par le format de l'interface. Une autre source est la base très riche des Classiques Garnier, mais là aussi, nous sommes limités par l'interface et par le fait qu'il n'est pas possible de télécharger des textes. Dans le cas de Frantext, la question des droits est connue, dans l'autre cas, il est dommage que ces textes, souvent créés par des chercheurs membres de la fonction publique, soient uniquement sous le contrôle d'un éditeur privé. La seule solution, et la solution logique, est qu'au lieu de se faire exproprier, les chercheurs français partagent leurs textes en les rendant disponibles librement sous licence non-commerciale *creative commons*. C'est ce que nous pratiquons.

Je donnerai dans ce qui suit un aperçu de deux projets, afin de montrer l'emploi des outils méthodologiques des HN : la numérisation de l'œuvre de Louis de Boissy (1694-1758) et sa mise dans son environnement, en commençant par celui du théâtre, et la création d'une édition numérique en libre accès de l'édition de 1701, révisée et augmentée par Basnage de Beauval, du dictionnaire de l'Abbé Furetière. La numérisation du premier est en train d'être réalisée en utilisant une collection de pièces de théâtre publiée par Néaulme en 1758 qui est la propriété de M^{me} Galleron. La source pour le dictionnaire de Furetière est un document PDF créé et hébergé par Gallica. Le projet Boissy est soutenu par le consortium CAHIER de la TGIR HumaNum ; le

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

projet Furetière sera réalisé par *crowdsourcing* et est toujours en phase exploratoire. Bien que le projet Boissy soit le plus avancé, il est intéressant méthodologiquement de commencer par le dictionnaire de Furetière.

8 Le projet Furetière

Le dictionnaire de Furetière est vite devenu une œuvre incontournable au XVIII[e] siècle, malgré l'opposition et la concurrence de l'Académie Française. L'histoire de ce dictionnaire a été admirablement décrite par Alain Rey (2006) donc nous n'entrerons pas dans les détails ici. Une version numérique existe chez les Éditions Garnier, présentant le désavantage d'une interface qui empêche l'accès au fichier XML et bloque donc la possibilité de faire une vraie analyse de texte en comparaison avec d'autres textes. Une tentative de numérisation a été réalisée par Chantal Wionet et Agnès Tutin (2001), mais en utilisant le SGML. Elles ont uniquement transcrit la lettre D, et leurs données ne sont pas disponibles.

Notre objectif est de fournir l'ensemble de la version de 1701 en consultation libre, mais aussi de lier le dictionnaire à un corpus de textes du XVII[e] siècle afin de voir l'environnement langagier au moment de sa création. Le dictionnaire de Furetière est une œuvre descriptive de la langue française de l'époque et est extrêmement riche en données phraséologiques. Notre objet donc est à la fois de faire la cartographie des idiomes, et notamment des collocations, de lier les citations aux textes cités, de préférence en version numérique, et d'analyser le langage de certains domaines thématiques. Puisque le dictionnaire est de taille importante, et que les difficultés de codage sont non négligeables, nous travaillerons dans un premier temps sur les mots liés au théâtre, afin de lier les deux projets Boissy et Furetière et de cartographier les mots annotés comme étant des termes.

Le dictionnaire est librement disponible en tant que PDF. Malgré les avancées dans la technologie OCR, il est très difficile, voire impossible, de numériser un tel texte, aussi bien à cause de la qualité de l'impression d'origine qu'à cause de l'image PDF. Il y a donc deux solutions, soit d'imprimer le PDF et d'utiliser un stabilo pour annoter le texte, soit d'utiliser un outil qui permet d'annoter le PDF. La première solution est la meilleure. Cependant, il est impossible de faire un travail totalement rigoureux avec cette technique, vu la quantité de données. L'alternative est d'utiliser un CAQDAS qui permet d'annoter et de coder le texte.

Les PDF sont chargés dans la mémoire d'Atlas.Ti, qui permet de mettre en valeur les passages à analyser. On peut dire qu'il s'agit d'un stabilo numérique.

Quand on sélectionne un passage (fig. 1), le morceau de texte est enregistré en tant que citation (*quotation*). On peut nommer la citation, nous le faisons avec le nom de l'entrée dictionnaire. Mais le plus intéressant est de pouvoir coder les éléments. Le code est libre et récursif. Au fil de la lecture, d'autres éléments importants deviennent visibles, et donc le codage est augmenté. Chaque code est lié à une ou plusieurs citations, ce qui permet de naviguer dans un texte et de regrouper les informations importantes. Dans cet exemple, nous codons les entrées marquées par Furetière comme étant des termes. Pour cela, il a utilisé plusieurs formules comme « Terme de droit »

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

ou « X est aussi un terme de [domaine] ». Prenant le cas de l'entrée pour « affier », nous notons qu'il s'agit d'un terme, par opposition à un mot de la langue générale, que le terme est annoté avec la formule « terme de [domaine] » : il s'agit donc d'un terme de domaine par opposition à un terme de métier. On voit enfin que le terme est lié à l'agriculture. Ainsi, nous avons pu lister l'ensemble des domaines et des métiers illustrés par Furetière et Basnage, et nous avons également noté les différentes formules employées pour décrire ces termes.

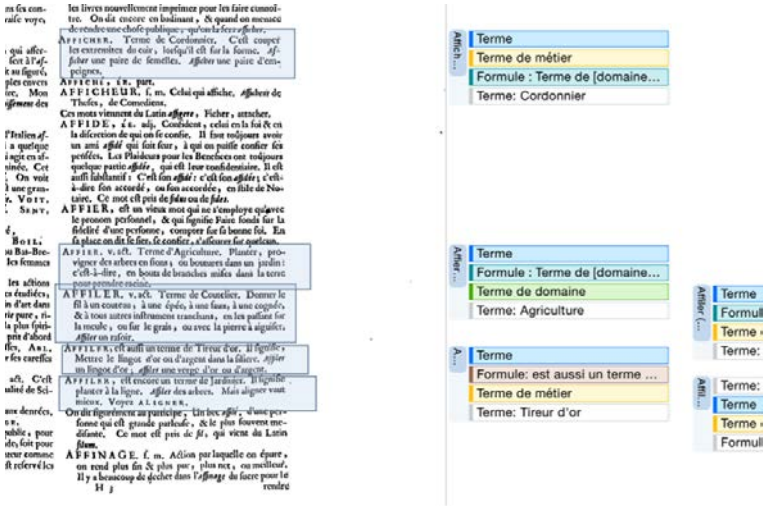


Figure 1 – Codage des termes

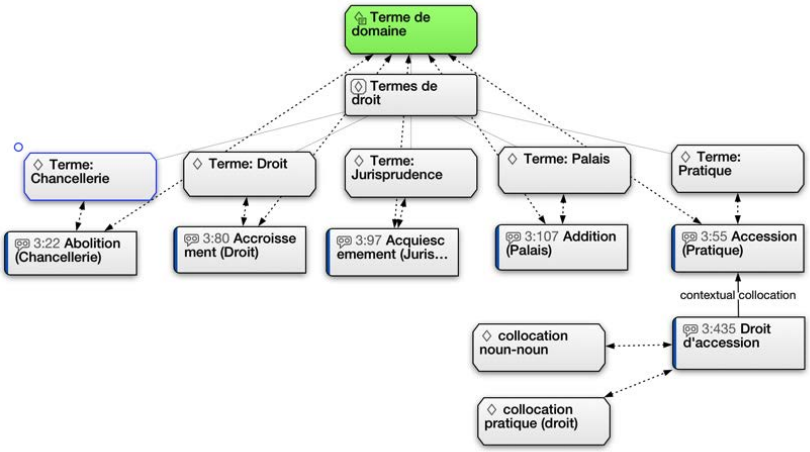


Figure 2 – Les réseaux de termes en droit

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

Une autre possibilité est de pouvoir classer et organiser les contenus. Dans le cas des termes juridiques, par exemple, plusieurs domaines sont associés. En utilisant la fonction réseau (fig. 2), nous pouvons monter les relations entre domaines et termes. Chaque lien est évidemment cliquable afin de pouvoir visualiser les termes en contexte.

Furetière n'était pas totalement systématique. La création des réseaux et l'analyse des contenus nous permet de trouver et d'énumérer d'autres termes afin d'avoir une vision complète d'un domaine dans le dictionnaire.

Il va de soi que les contenus phraséologiques nous intéressent, donc là aussi nous faisons un codage afin d'isoler les différents types d'idiomes. Furetière était fasciné par la marine, nous avons non seulement des termes, mais des collocations décrivant les différentes actions à bord un bateau (fig. 3). Nous codons à la fois le fait qu'il s'agit d'une collocation, d'un exemple, d'un proverbe, etc., et la nature de l'élément, par exemple une collocation verbe – nom.

un droit qu'on payoit en especes, & dont le prix étoit incertain. Il eſt *abonné* à tant par an pour tous droits Seigneuriaux. Ce Marchand eſt *abonné* à cent écus par an avec le Doüanier, pour les droits d'entrée de toutes ſes marchandiſes. Il ſe dit avec le pronom perſonnel : Je m'*abonnai*, je me ſuis *abonné*. Par pluſieurs Coutumes les roucins de ſervice ſont *abonnez* à un écu.

ABONNER, ſignifie auſſi, Aliener, changer : c'eſt quand un vaſſal aliene ſes rentes, ou change ſon hommage à quelque autre devoir. Voyez les Coutumes d'Anjou & du Maine. L'ancienne Coutume de Tours portoit *abiener*, au lieu d'*abonner*, qui eſt en la nouvelle.

ABONNÉ, ÉE. part. paſſ. & adj. Champart *abonné*, ou *abourné*. Les Coutumes font auſſi ſouvent mention d'hommes & de femmes ſerfs *abonnez*, de quête *abonnée*, d'aides *abonnées*, c'eſt-à-dire fixées.

On dit auſſi des Meuniers *abonnez* au Seigneur, pour avoir permiſſion de chaſſer, & de chercher les mounées dans ſa Seigneurie.

On dit auſſi, Taille *abonnée* en la Coutume de Nevers, & *abournée* en la Coutume de Troyes.

ABONNIR. v. act. Rendre, devenir meilleur. Les cabaretiers trouvent moyen d'*abonnir* leur vin par les drogues qu'ils y mêlent. On le dit auſſi avec le pronom perſonnel. Cet homme s'*abonnit* tous les jours depuis qu'il hante les gens de bien. Les affaires criminelles s'*abonniſſent*, quand on les fait tirer en longueur. Les fruits s'*abonniſſent* en meuriſſant. Ce mot ſe tire du Latin *bonus*, bon. Il ne ſort point de la converſation.

ABONNIR, terme de Potier, ſignifie, Faire ſecher le carreau, & le mettre en état de rebattre.

ABORD. ſ. m. Entrée, accés, approche. Les *abords* de cette place ſont dangereux. Toutes les côtes d'Angleterre & de Hollande ſont de difficile *abord*. Le commerce fleurit d'ordinaire dans les ports qui ſont de facile

ABORDABLE. adj. m. & f. Acceſſible, accoſtable. Cette côte n'eſt pas *abordable* à cauſe des écueils. Cet homme eſt ſi glorieux qu'il eſt *abordable* à peu de perſonnes.

ABORDAGE. ſ. m. Terme de marine, qui ſe dit lors que deux vaiſſeaux ſe heurtent, ou s'accrochent pour ſe combattre. Aller à l'*abordage*, ſe dit de l'action d'un vaiſſeau qui en a joint un autre pour l'enlever. Faire l'*abordage* en belle, ou debout au corps, c'eſt-à-dire, l'éperon dans le flanc. L'*abordage* de franc étable eſt celui qui ſe fait par le devant, & en droiture, pour s'enferrer par les éperons. L'*abordage* de bout au corps, c'eſt mettre l'éperon dans le flanc.

ABORDAGE, ſe dit auſſi du choc de deux vaiſſeaux du même party, ſoit lors qu'ils vont en flotte, ſoit lors qu'ils ſont en même mouillage; ce qui arrive par la violence des flots ou des vents qui les portent, & qui les font deriver les uns ſur les autres.

ABORDER. v. act. & n. Arriver en quelque lieu, ſpecialement par mer; prendre terre. J'*aborde*, j'*abordai*, je ſuis *abordé*. La flotte des Indes eſt *abordée* en Eſpagne. Les Marchands *abordent* de tous côtez à la foire de Beaucaire le 21. de Juillet. On ne convient pas qu'Enée ſoit *abordé* en Italie. SENT. DE CL. Il n'eſt pas ſeur d'*aborder* de cette côte, parce que la mer ſe retirant, les vaiſſeaux y demeurent à ſec. ABL. Il ne put *aborder* à cauſe que la rive étoit eſcarpée. ID. Ils *aborderent* en des païs inconnus. VAU. Il ſignifie, Arriver en foule. Les preſens *abordent* chez moi de toutes parts. ABL. Il ſignifie encore, Entrer, parvenir. Nous ne pumes *aborder* de la place, parceque toutes les avenuës étoient gardées. Il fut impoſſible d'*aborder* juſqu'à l'autel à cauſe de la foule du peuple.

ABORDER, ſignifie auſſi, Venir à bord d'un vaiſſeau. On a contraint ce vaiſſeau ennemi de mettre pavillon bas, & d'*aborder*. On dit de deux vaiſſeaux qui s'ap-

Figure 3 – Les collocations d'abordage

Nous sommes en train de travailler sur la deuxième édition du dictionnaire, il est donc intéressant de comparer les versions de 1690 et de 1701, et aussi le dictionnaire du langage scientifique de Corneille de 1694, ouvrage commandé en urgence par l'Académie afin de contrer le succès du dictionnaire de Furetière.

Les possibilités offertes par un bon outil CAQDAS sont énormes et offrent beaucoup d'avantages pour une analyse complète ou une étude de faisabilité. Il s'agit d'outils

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

créés pour les sciences humaines, mais il n'y pas de raison que, dans le cadre des HN, de tels outils ne soient pas adoptés. Il est intéressant de noter que beaucoup de CAQDAS intègrent un concordancier – des outils de la linguistique de corpus au service de la sociologie.

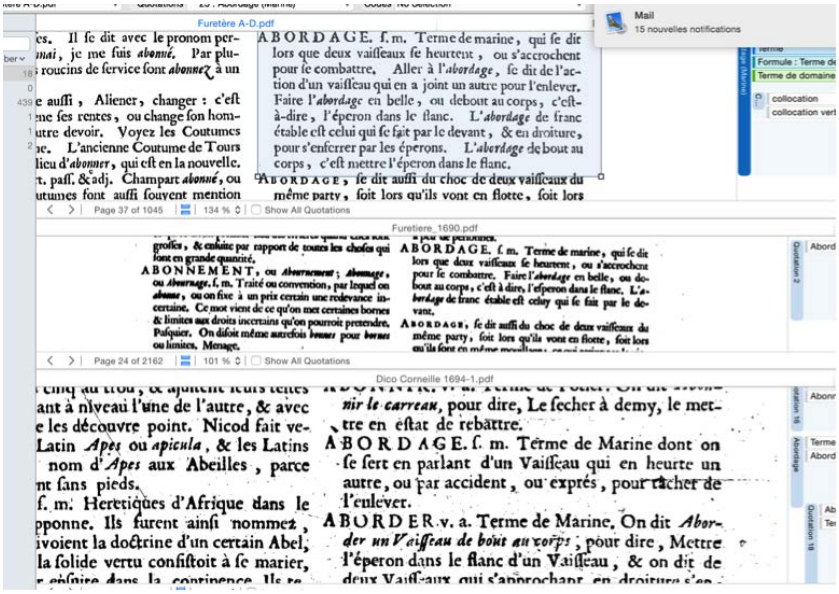


Figure 4 – Trois dictionnaires : Furetière 1690, Furetière-Basnage 1701 et Corneille 1694

À terme, nous voulons mettre l'ensemble du dictionnaire en ligne, mais avant de le faire, une analyse par CAQDAS nous permet de cartographier les points clé, de repérer des difficultés et de fournir des résultats intermédiaires avant de nous attaquer à une numérisation complète. L'analyse des réseaux thématiques nous permet de travailler dans un premier temps uniquement sur les domaines choisis – le droit, la marine et le théâtre – mais aussi de faire le travail de recontextualisation grâce au lien avec des textes employés par Furetière et Basnage pour illustrer leurs dires. L'étape suivante est la saisie et le balisage des entrées thématiques selon les recommandations de la TEI.

La TEI permet de fournir un texte accessible à tous et qui peut être analysé avec des outils différents et même traité en tant que base de données. Pour illustrer les possibilités offertes par la TEI, le mieux est de jeter un coup d'œil sur le projet Boissy.

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

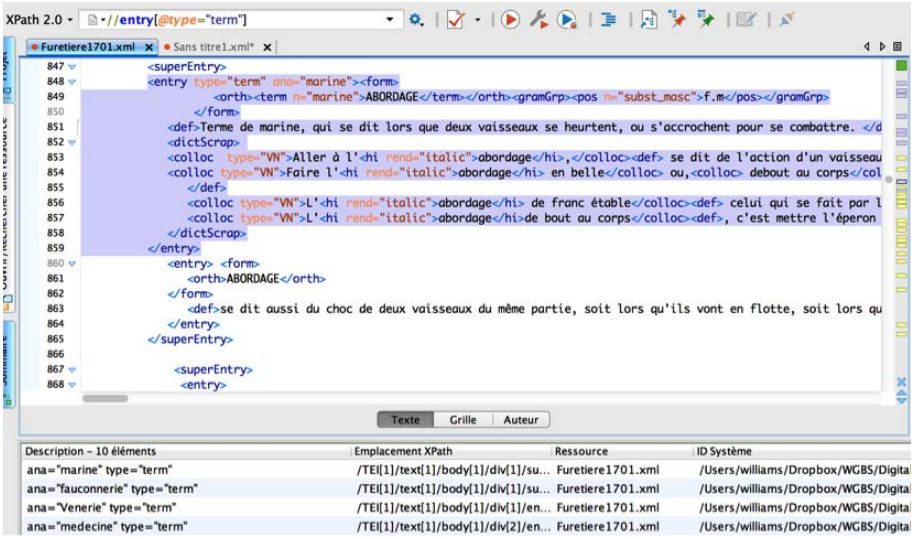


Figure 5 – Le terme « abordage » en TEI

9 Le projet Boissy

Le point de départ du projet Boissy a été une analyse des contenus comme pour Furetière. La différence est que nous sommes partis d'un ouvrage imprimé, ce qui nous a permis de commencer tout de suite à saisir les textes pièce par pièce. À ce jour, nous avons mis en ligne quatorze pièces entièrement balisées en XML selon les recommandations de la TEI. La plupart des textes sont disponibles dans deux versions, une version XML et une version HTML obtenue par une transformation XSLT et corrigée pour la rendre plus lisible à l'écran. Nous associons des feuilles de style à chaque pièce, ce qui permet pour les textes HTML de lire soit l'orthographe originelle, soit l'orthographe moderne. Il est probable que par la suite nous ne mettrons que les versions XML avec une feuille de style améliorée.

Pour les pièces de Boissy, nous avons rempli une des missions des HN qui est de rendre librement disponibles des sources textuelles. Cependant, l'objet n'est pas simplement de pouvoir visualiser les textes, mais aussi de les interroger. Les balises XML, et plus encore, les attributs, nous permettent de faire des requêtes très poussées afin d'extraire des informations pertinentes pour nos propres recherches. Comme toujours avec l'informatique, il y a une part de mystère soigneusement entretenue, mais non seulement la TEI n'est pas si difficile à apprendre, mais une fois une structure réalisée en collaboration, il est possible d'affiner le balisage, et donc les recherches possibles.

Pour mémoire, un document TEI est composé de deux grandes parties ; l'en-tête ou *header* qui regroupe les informations bibliographiques et de codage, et le texte. Le codage consiste en trois groupes d'éléments : les balises, les attributs et les entités. Dans ce texte, nous ne parlerons pas des entités, sauf pour dire que le codage d'un caractère, comme le s long, est géré par des entités « caractères ». Dans la figure 6, nous pouvons voir comment une pièce peut être codée :

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

```
</front>
<body>
    <pb n="253"/>
    <div1 type="title">
        <head>LE RIVAL <lb/> FAVORABLE,<lb/> COMEDIE.</head>
    </div1>
    <div1 type="play">
        <div2 type="acte" n="1">
            <head>ACTE PREMIER.</head><lb/>
            <div3 type="scene" n="1">
                <head>SCENE PREMIERE.</head><lb/>
                <stage>LA FLEUR, MARTON.</stage><lb/>
                <sp who="La_Fleur_RF">
                    <speaker>LA FLEUR.</speaker><lb/>
                    <l><c type="lettrine">J</c>e viens t'interroger fur le fort de mon <lb/>Maître ;</l>
                    <l part="I"><name type="person" n="Clarice">Clarice</name> l'aime-t'elle aujourd'hui ?</l>
                </sp>
                <sp who="Marton_RF">
                    <speaker>MARTON.</speaker><lb/>
                    <l part="F">Mais, peut-être,</l>
                    <l>Je ne puis, fur ce point, te parler nettement.</l>
                </sp>
                <sp who="La_Fleur_RF">
                    <speaker>LA FLEUR.</speaker><lb/>
                    <l>Dans une <choice><orig>ame</orig><reg>âme</reg></choice> coquette on lit <choice><orig>mal-aifément</orig>
                </sp>
                <sp who="Marton_RF">
                    <speaker>MARTON.</speaker><lb/>
                    <l>Non, celle de <name type="person" n="Clarice">Clarice</name> eſt <choice><orig>plûtôt</orig><reg>plutôt</r
                    <l>C'eſt l'inégalité qui la caractérife,</l>
                    <pb n="254"/>
                    <l>Son cœur, depuis qu'il eſt combattu par l'amour,</l>
                    <l>Semble en <choice><orig>bifarrerie</orig><reg>bizarrerie</reg></choice> augmenter chaque jour.</l>
```

Figure 6 – Le codage TEI dans une pièce de théâtre

Nous pouvons noter que la pièce est divisée en mais que chaque division est nommée en tant que pièce (*play*), acte, scène. À l'intérieur de la balise, la typologie de la division est le numéro de l'acte ou scène. Il s'agit donc d'attributs. Parmi les balises importantes se trouve la balise <sp>, qui correspond au tour de parole. À l'intérieur, nous pouvons coder le locuteur, <speaker>. Ainsi une prise de parole dans une pièce de théâtre prend la forme suivante :

<sp who="#Marton">
<speaker>MARTON.</speaker><lb/>
<l>Mais, peut-être,</l>
<l>Je ne puis, fur ce point, te parler nettement.</l>
</sp>

Dans ce cas, nous savons que c'est MARTON qui prononce un certain nombre de lignes. Le nom MARTON apparaîtra dans une version en ligne et dans une version papier. Cependant, à l'intérieur de la balise nous avons l'attribut *who* qui nous permet de dire à la machine qui parle et ainsi d'extraire l'ensemble des paroles de Marton. Il est possible d'ajouter de nombreux attributs afin de différencier qui parle et à qui on s'adresse, les différent rôles des personnages, les dates, les lieux, etc. La liste est sans fin et au-delà des informations et balises de base on peut l'adapter en fonction des besoins des chercheurs. On peut dire que le texte est parlant pour un être humain, mais que les attributs parlent à la machine. Ainsi, les balises et les attributs nous permettent d'utiliser les pièces comme une base de données, et de l'interroger avec des outils comme BaseX.

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

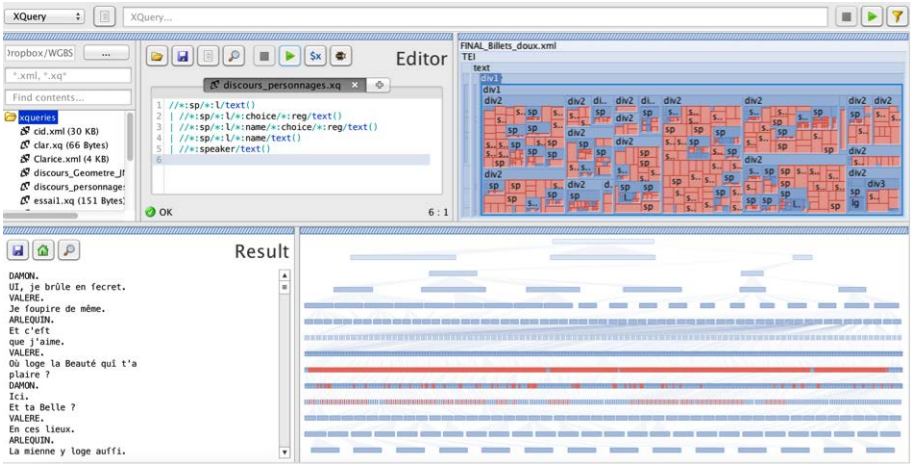


Figure 7 – Une extraction avec BaseX

Certes, BaseX est difficile d'emploi pour un non-informaticien, mais il est possible de faire écrire une série de requêtes selon nos besoins et puis de les appliquer nous mêmes par la suite. Il faut aussi regarder la question différemment. BaseX est difficile puisque créé par et pour les spécialistes du XML et des bases de données, mais une fois le système apprivoisé selon nos besoins, il devient possible de faire ou de faire faire une interface plus conviviale. Cependant, sans le balisage XML et une série d'essais sur un outil comme BaseX, il est impossible de demander la création d'un outil ou d'une interface. La collaboration en connaissance de cause entre chercheur humaniste et informaticien est essentielle.

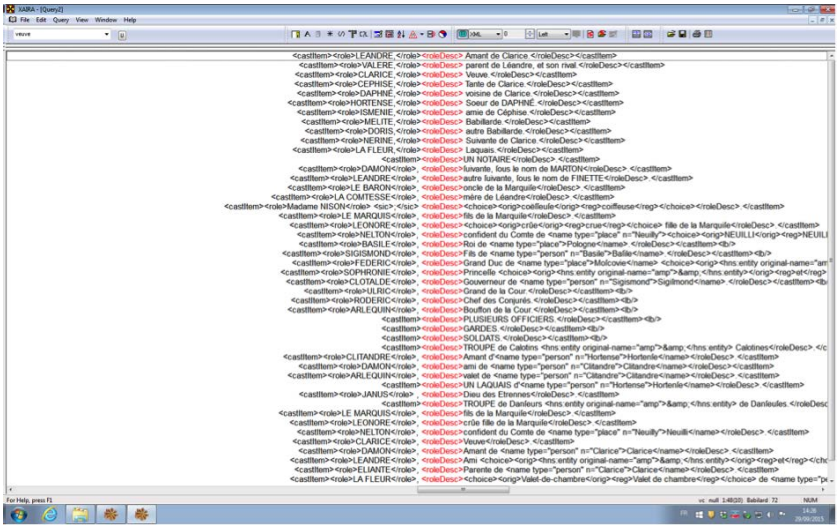


Figure 8 – XAIRA comme concordancier XML

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

Dans le monde de la linguistique de corpus, les outils qui sont capables de lire le XML existent déjà, par exemple XAIRA, l'outil crée pour le British National Corpus, qui est disponible gratuitement [6]. Puisque XAIRA est issu de la linguistique de corpus, il est possible de l'utiliser comme un concordancier simple, mais comme nous le montrons dans la figure 8, il est également possible de faire des recherches dans le XML.

10 Conclusion

S'il n'y a pas de sens hors contexte, il faut chercher le contexte, et les outils des humanités numériques nous permettent de le faire. Il existe des outils qui sont facilement utilisables en l'état, comme les CAQDAS, mais qui sont simplement inconnus des chercheurs des sciences humaines. Dans ce cas, le rôle des HN est de partager les connaissances. Il y a des outils pour lesquels il faut faire un effort pour apprendre un minimum de leur structure informatique, c'est le cas avec la TEI. Certes, la TEI est un peu plus difficile d'accès, mais travailler en HN ne signifie pas simplement prendre ce que l'informatique nous propose, mais faire un bout de chemin vers l'autre, collaborer. Il y a aussi des outils qui demandent de comprendre un minimum sur l'origine et la méthodologie d'une autre discipline, c'est le cas de la linguistique de corpus/*corpus linguistics*, qui est déjà une discipline humaniste.

Si les humanités numériques sont une discipline jeune, il faut les créer à notre image en adoptant des outils existants, en collaborant avec les informaticiens en connaissance de cause afin de créer d'autres outils. Il faut créer et ouvrir des ressources libres puisque le contexte historique et culturel se trouvera dans les ressources partagées. Le contexte est dans les textes, le partage fait la force.

Bibliographie

Brugidou, Mathieu, Le Roux, Dominique (2005) « L'usage des CAQDAS et des logiciels d'analyse textuelle dans l'analyse secondaire », Colloque « L'analyse secondaire en recherche qualitative » – Grenoble. Disponible sur http://archipolis.hypotheses.org/files/2014/06/Lusage-des-CAQDAS-et-des-logiciels-danalyse-textuelle-dans-lanalyse-secondaire.pdf [consulté le 20 juin 2020].

Engwall, Gunnar (1994) « Not Chance, but Choice : Criteria in Corpus Creation », dans *Computational Approaches to the Lexicon*, Atkins S., et Zampolli A. (dir.)., Oxford, Clarendon Press, p. 49-82.

Firth, John Rupert (1957) *Papers in Linguistics 1934-1954*, Oxford, Oxford University Press.

Galleron, Ioana (2015) « Pour comprendre le sacré : la scénographie de la chute des anges », dans *Le Sacré en question. Bible et mythes sur les scènes du XVIII^e siècle*, Ferrier, B. (dir.), Paris, Classiques Garnier, p. 119-138.

Hanks, Patrick (2013) *Lexical Analysis : Norms and Exploitations*, Massachusetts, MIT Press.

Hanks, Patrick (1986) « Typicality and meaning potentials », *ZüriLEX'86 Proceedings*, Zurich, Francke Verlag, p. 37-47.

Herrenschmidt, Clarisse (2007) *Les trois écritures : langue, nombre, code*, Paris, Éditions Gallimard.

Hoey, Michael (2005) *Lexical Priming : A New Theory of Words and Language*, London, Routledge.

Krishnamurthy, Ramesh (dir.) (2004) *English Collocation Studies : The OSTI Report*, London, Continuum.

Leavy, Patricia (dir.) (2014) *Oxford Handbook of Qualitative Research*, Oxford, Oxford University Press.

6. http://xaira.sourceforge.net [consulté le 13 novembre 2018].

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843

Louw, William (1993) « Irony in the text or insincerity in the writer ? The diagnostic potential of semantic prosodies », *Text and Technology*, Baker M. (dir.), Amsterdam, John Benjamins.

Rey, Alain (2006) *Furetière : un précurseur des Lumières sous Louis XIV*, Paris, Fayard.

Sinclair, John (2005) « Corpus and Text : Basic Principles », *Developing Linguistic Corpora : A Guide to Good Practice*, Wynne M. (dir.), Oxford, AHDS, p. 1-16.

Sinclair, John (1991) *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

Sinclair, John (dir.) (1987) *Looking Up : an Account of the COBUILD Project in Lexical Computing*, London, Collins.

Sinclair, John (2004) *Trust the Text : Language, Corpus and Discourse*, London, Routledge.

Teubert, Wolfgang (2010) *Meaning Discourse and Society*, Cambridge, Cambridge University Press.

Tognini-Bonelli, Elena (2001) *Corpus Linguistics at Work*, Amsterdam, John Benjamins.

Wionet, Chantal et Tutin, Agnès (2001) *Pour informatiser le Dictionnaire universel de Basnage (1702) et de Trévoux (1704). Approche théorique et pratique*, Paris, Honoré Champion.

Williams, Geoffrey (2006) « La linguistique et le corpus : Une affaire prépositionnelle », *Texto, revue de linguistique en ligne*. http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Williams.pdf [consulté le 19 juin 2020].

Williams, Geoffrey (2010) « Many rooms with corpora », *International Journal of Corpus Linguistics,* 15 :3, p. 400-408.

Published in : Meynard C. ; Lebarbé T. ; Costa S. (dir.) (2020), *Patrimoine et Humanités numériques*, Editions des archives contemporaines, ISBN : 9782813003843, doi : https ://doi.org/10.17184/eac.9782813003843