Le rôle des métadonnées

De l'accessibilité des sources à l'élaboration des objets de la recherche (données et modèles)

Marie-Hélène LAY

Université de Poitiers FoReLLIS (EA 3816)

1 Introduction : de la disponibilité à l'utilisation des données

La communauté des sciences humaines, et tout particulièrement celle du consortium CAHIER, se réjouit du développement des *humanités numériques* et de la disponibilité croissante de données, qu'elles soient textuelles, orales, ou multi-modales.

L'enthousiasme éprouvé à l'idée de disposer de tant de sources et ressources se trouve cependant tempéré par deux caractéristiques constitutives du paradigme numérique. D'une part, on prend conscience du fait qu'il ne suffit pas de disposer des textes « saisis », par exemple en format .txt, .odt ou .doc, pour en faire des données de la recherche exploitables par/pour la communauté ; d'autre part, on se confronte à un phénomène parfois appelé « infobésité » : la masse de données est telle qu'on ne sait pas toujours comment en faire bon usage, bon usage scientifique s'entend. Tout ceci suppose que les données de la recherche aient été patiemment élaborées pour cerner au plus près les éléments permettant de nourrir réflexion et argumentation autour d'une question posée. Ce qui est nouveau dans l'environnement numérique, c'est que l'on peut apporter une réponse unifiée à toutes ces problématiques grâce à la pratique de l'annotation, de l'adjonction de métadonnées ; celle-ci permet de structurer les fonds documentaires, de localiser les ressources et de documenter les perspectives scientifiques au sein même des textes.

L'annotation du document et de son contenu est une opération extrêmement coûteuse en moyens humains. De ce fait, il peut arriver que les chercheurs négligent cette étape, pour aller plus vite vers leur étude, l'annotation est donc réduite au stade de préparation des données. Mais cette réduction de l'annotation à une étape préalable au travail

de recherche en soi est au coeur du problème de la sous-utilisation des corpus annotés, comme nous le verrons plus loin : la conséquence est que le chercheur accédant aux ressources textuelles au format numérique reste dans une approche pré-numérique des textes.
L'*annotation de contenu*, ou *tagging*, ne sera pas abordée ici en tant que pratique unifiée permettant la réutilisabilité des données – ce qui est le point de vue adopté pour la mise à disposition de grands corpus dits « corpus de référence », comme la base *Frantext*[1] ou le *Corpus of Contemporary American English* (*COCA*)[2] par exemple –, ni comme une étape « utile » avant le recours à des outils de traitement statistique, par exemple. Nous n'aborderons donc pas cette question par le biais des standards d'annotation tels qu'ils sont abordés par la très active communauté de la *Text Encoding Initiative*. L'*annotation manuelle de contenu* sera réinscrite ici dans la lignée des pratiques pré-numériques de l'enrichissement des textes, enrichissement au sens où une information nouvelle est associée à une information initiale. C'est ce que nous faisons communément en préparant une lecture critique, en produisant diverses versions d'un texte, en corrigeant des épreuves pour un éditeur, en corrigeant des travaux d'étudiants. Dans les pratiques d'annotation manuelle dont il est question ici, les métadonnées sont généralement insérées dans le document lui-même[3].

Annoter, est, formulé de la façon la plus simple, ajouter des « notes » à un document, ajouter une information B à une information initiale A ; enrichir des données (information A) par d'autres données (information B) appelées métadonnées. En des termes plus techniques, on peut emprunter à Kären Fort (2012 : p. 17) la définition globale suivante :

> l'annotation recouvre à la fois le processus consistant à apposer (ad-) une note sur un support, l'ensemble des notes ou chaque note particulière qui en résulte et ce, sans préjuger *a priori* de la nature du support considéré (texte, vidéo, images, etc.), du contenu sémantique de la note (note chiffrée, valeur choisie dans un référentiel fermé ou texte libre), de son positionnement global ou local, ni de son objectif (visée évaluative ou caractérisante, simple commentaire discursif).

Notre objectif est ici de montrer en quoi la pratique de l'annotation manuelle n'est pas une contrainte supplémentaire imposée par la mise à disposition des informations au format numérique, mais un outil de façonnage des données de la recherche, un outil d'expérimentation des modèles que nous élaborons, un laboratoire de formulation et de test, un outil pertinent venant à l'appui de nos heuristiques[4]. Pour ce faire, les exemples proposés donneront à voir comment des pratiques qui sont déjà les nôtres peuvent aisément embrasser un paradigme réputé plus « scientifique », celui de la

[1] www.atilf.fr/ressources/frantext/, consulté le 10 septembre 2021.
[2] https://www.english-corpora.org/coca, consulté le 10 septembre 2021.
[3] On peut aussi faire de l'annotation dite « déportée », c'est-à-dire qu'elle donne lieu à un fichier séparé du corpus originel, mais synchronisé avec celui-ci. Entrer dans les détails de cette technique n'apporterait rien ici à notre propos.
[4] Le Moigne (1991) donne la définition suivante : « Une heuristique est un raisonnement formalisé de résolution de problème (représentable par une computation connue) dont on tient pour plausible, mais non pour certain, qu'il conduira à la détermination d'une solution satisfaisante du problème. » Pour une présentation concrète d'une formalisation algorithmique d'une boucle heuristique d'analyse de texte, voir Lay (2015).

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

démarche expérimentale empirique de type *bottom up* (des données vers les modèles) en répondant aux critères de reproductibilité de la recherche, de réplicabilité et de possibilité de falsifier des résultats, critères fondamentaux retenus par la recherche scientifique et par les évaluateurs institutionnels des projets que nous déposons.

2 Contribution de l'annotation à l'activité de recherche

Quel que soit l'environnement dont on dispose, l'annotation est une tâche coûteuse en temps, il est donc tout à fait légitime de se demander à quoi elle sert et en quoi elle s'inscrit dans le processus même de l'activité de recherche.

Pour clarifier la place que peut prendre, dans nos disciplines, l'annotation manuelle de contenu dans les environnements numériques, il nous semble nécessaire d'interroger celle qu'elle occupe sous sa forme traditionnelle. Nous pourrons ainsi en examiner la portée et établir dans quelle mesure son outillage informatique relève de la simple transposition ou induit un changement de paradigme comparable aux mutations observables par ailleurs dans le champ documentaire, la plasticité des documents numériques ayant profondément modifié la définition de ce que sont les textes et les documents.

Sous sa forme traditionnelle, l'annotation est une pratique répandue en dehors même de l'activité scientifique d'étude des données disponibles, elle consiste à enrichir un texte d'informations diverses, avec un objectif déterminé. De ce strict point de vue, on peut par exemple comparer les annotations de Montaigne sur les *Essais*[5] à d'autres annotations visant à l'amendement d'un texte, que ce soit lors de la finalisation d'un « bon à tirer » ou lors de la correction de copies. Après avoir observé les points communs de ces pratiques, nous verrons comment la rationalisation et la systématisation des annotations correspond à *l'élaboration de catégories qui dépendent d'un point de vue et d'une tâche*. En effet, cette systématisation s'accompagne rapidement d'un référentiel stable de désignation : dès lors, les métadonnées de contenu sont des métadonnées comme les autres et les textes annotés se transforment en objet de la recherche *permettant d'élaborer des modèles et d'en observer les manifestations*. Pour mener ces tâches à bien, de nouveau, les outils numériques s'inspirent de pratiques anciennes, celles de l'établissement d'index et de concordances (on y reviendra), destinés à discerner, dans la profusion des observables, ce qui est « identique » (et relève de la même catégorie) de ce qui est différent. De fait, une telle discrimination ne se pense pas « hors contexte », et les modélisations sont des opérations finalisées en fonction de leurs objectifs. Il est fondamental de bien en saisir la portée au niveau de l'activité modélisante en soi, et pas de sa seule mise en œuvre (ce que l'on peut être tenté de faire lorsqu'on aborde la question des outils). Ce point de vue rejoint la position énoncée par André Martinet (1979) sur la détermination de catégories grammaticales :

> Toute description suppose une sélection. […] Une description est nécessairement finie, ce qui veut dire que seuls certains traits de l'objet à décrire pourront être dégagés. Ceux que relèvent deux personnes différentes ont toutes les chances de

[5] https://montaigne.univ-tours.fr/lexemplaire-de-bordeaux-des-essais-de-montaigne-numerise-et-en-ligne/, consulté le 10 septembre 2021. Voir la vidéo disponible sur cette page.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

ne pas être les mêmes[6]. **Toute description sera acceptable** à condition qu'elle soit cohérente, **c'est-à-dire qu'elle soit faite d'un point de vue déterminé**[7]. Une fois ce point de vue adopté, certains traits, dits pertinents, sont à retenir.

Ce qui est vrai pour les catégories grammaticales et la consistance du discours que les grammaires construisent sur le fonctionnement d'une langue l'est pour tous les discours que nous construisons sur nos objets de recherche.

Ancrage dans des pratiques anciennes sous-jacentes aux opérations de modélisation, usage au cœur même des processus heuristiques, inscription dans les allers-retours *bottom-up* et *top-down* de l'élaboration des métadonnées, tels sont donc les axes que nous retiendrons ici pour interroger la place de l'annotation manuelle.

3 Une pratique ancienne et répandue : la correction d'un document

Afin de formuler les caractéristiques de ce que sont les « notes », nous proposons ici une page de Montaigne (Fig. 1), annotée entre 1588 et 1592[8].

On voit ici ce dont il est question : un texte initial a été enrichi de diverses façons. Il peut s'agir de corrections localisées d'ordre typographique, du remplacement d'un mot par un autre, de développements ou de commentaires plus longs. Une note peut être de longueur variable et porter sur des empans de texte de taille variable.

Quoi qu'il en soit, l'observation de ce manuscrit nous permet d'établir que l'annotation mobilise deux dimensions :

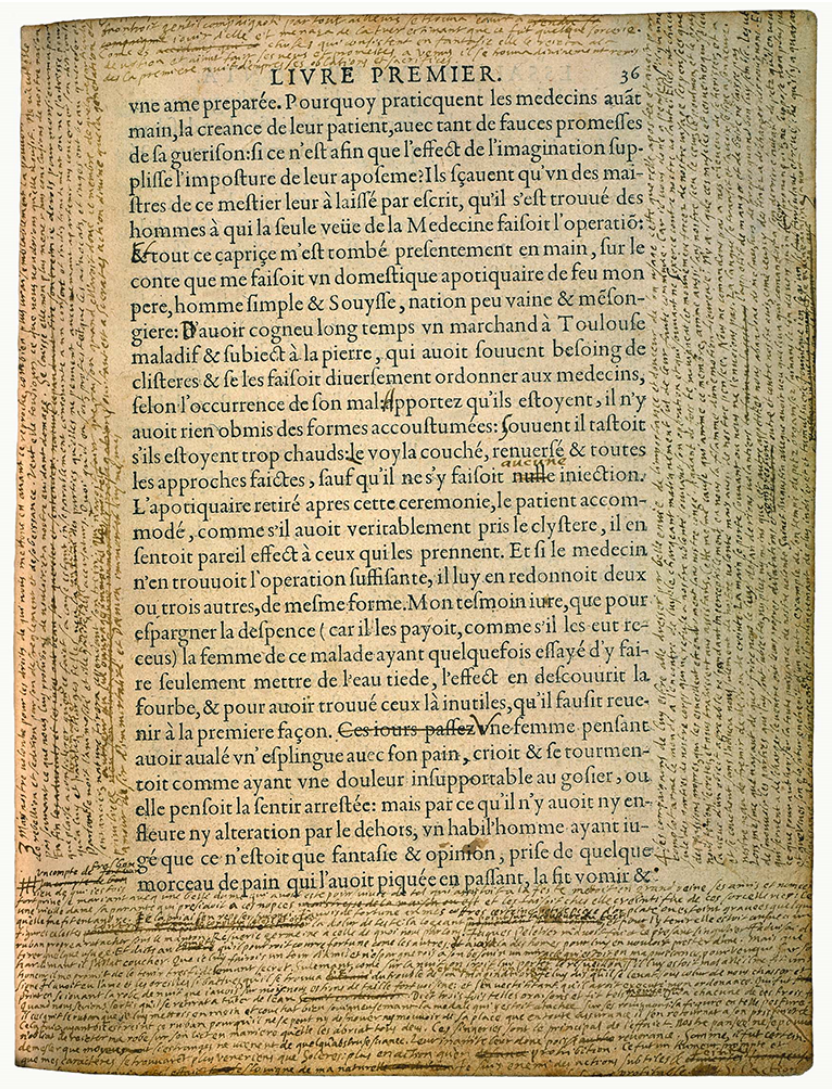
- la localisation (qui peut être, dans les faits, de divers types, par exemple : indexation, indication, renvoi) ;
- la dimension explicative élucidant le propos du marquage (l'information développée peut, elle aussi, être de divers types : valeur autorisée par un référentiel, commentaire, glose).

[6] On peut mentionner ici cet exemple issu d'une campagne d'annotation d'entités nommées structurées (Fort, 2012) : les choix divers faits pour l'annotation de l'expression « *territoires palestiniens* », qui peut être considérée comme une région, un pays ou une organisation administrative, selon le point de vue de l'annotateur (toutes ces étiquettes étant disponibles dans le modèle d'annotation). Pour s'en convaincre encore, si besoin est, on peut aussi consulter la description de *de* dans les différentes grammaires de référence du programme de l'agrégation de Lettres Modernes comme Damourette et Pichon (1911-1940), Riegel, Pellat et Rioul (1994), Sancier-Chateau et Denis (1997), Le Goffic (1993).

[7] Caractères gras dans le texte.

[8] Fac-similé de l'exemplaire de Bordeaux des Essais disponible sur Gallica et sur le site Internet des Bibliothèques Virtuelles Humanistes sur https://montaigne.univ-tours.fr/lexemplaire-de-bordeaux-des-essais-de-montaigne-numerise-et-en-ligne (consulté le 10 septembre 2021).

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352



LIVRE PREMIER. 36

vne ame preparée. Pourquoy praticquent les medecins auãt main, la creance de leur patient, auec tant de fauces promesses de sa gueriſon: ſi ce n'eſt afin que l'effect de l'imagination ſupplisse l'impoſture de leur apoſeme? Ils ſçauent qu'vn des maiſtres de ce meſtier leur à laiſſé par eſcrit, qu'il s'eſt trouué des hommes à qui la ſeule veüe de la Medecine faiſoit l'operatiõ: & tout ce caprice m'eſt tombé preſentement en main, ſur le conte que me faiſoit vn domeſtique apotiquaire de feu mon pere, homme ſimple & Souyſſe, nation peu vaine & mẽſongiere: D'auoir cogneu long temps vn marchand à Toulouſe maladif & ſubiect à la pierre, qui auoit ſouuent beſoing de cliſteres & ſe les faiſoit diuerſement ordonner aux medecins, ſelon l'occurrence de ſon mal: Apportez qu'ils eſtoyent, il n'y auoit rien obmis des formes accouſtumées: Souuent il taſtoit s'ils eſtoyent trop chauds: Le voyla couché, renuerſé & toutes les approches faictes, ſauf qu'il ne s'y faiſoit nulle aucune iniection. L'apotiquaire retiré apres cette ceremonie, le patient accommodé, comme s'il auoit veritablement pris le clyſtere, il en ſentoit pareil effect à ceux qui les prennent. Et ſi le medecin n'en trouuoit l'operation ſuffiſante, il luy en redonnoit deux ou trois autres, de meſme forme. Mon teſmoin iure, que pour eſpargner la deſpence (car il les payoit, comme s'il les eut receus) la femme de ce malade ayant quelquefois eſſayé d'y faire ſeulement mettre de l'eau tiede, l'effect en deſcouurit la fourbe, & pour auoir trouué ceux là inutiles, qu'il fauſit reuenir à la premiere façon. Ces iours paſſez vne femme penſant auoir aualé vn' eſplingue auec ſon pain, crioit & ſe tourmentoit comme ayant vne douleur inſupportable au goſier, ou elle penſoit la ſentir arreſtée: mais par ce qu'il n'y auoit ny enfleure ny alteration par le dehors, vn habil homme ayant iugé que ce n'eſtoit que fantaſie & opinion, priſe de quelque morceau de pain qui l'auoit piquée en paſſant, la fit vomir &

Fig. 1 : Annotations de Montaigne sur l'exemplaire de Bordeaux des *Essais*

L'annotation se manifeste donc chaque fois que l'on ajoute, *avec une intention particulière*, de l'information à de l'information (par exemple pour repérer, etc.) Le procédé retenu est une pratique aussi ancienne que la diffusion des textes écrits, dont les modalités de réalisation sont extrêmement variées comme en témoignent les exemples suivants à la base de la plupart de nos pratiques d'analyse :

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

- la préparation d'une traduction amènera à souligner des mots, isoler des tournures, colorier des éléments de progression, et ce *au service d'un texte cible* ;
- le « commentaire » de texte conduit à repérer une partie des unités lexicales, ou des caractéristiques de l'emploi des temps ou autres, et ce *au service d'une analyse* visant à mettre en évidence des points fondamentaux ;
- le résumé comme la sélection de citations imposent d'isoler les parties les plus *significatives d'un certain point de vue*, d'un trait de crayon ou d'un post-it marque-page ;
- l'activité lexicographique de rédaction des définitions repose sur le relevé d'un certain nombre d'éléments plus ou moins contextualisés, et ce *au service d'une certaine vision* du lexique portée par des enjeux sociétaux, etc.

La pratique étant très répandue et largement répétitive, une question émerge : celle de la « rationalisation » du procédé. On peut très souvent identifier des « notes récurrentes » que l'on va chercher à stabiliser ou *se doter d'un vocabulaire fixe* (recourir à tel signe pour noter tel type de correction à apporter[9], souligner les verbes en rouge, etc.), c'est-à-dire développer un standard d'annotation (même local). Ceci s'impose de façon d'autant plus évidente que l'annotation est (1) répétitive sur un même document (2) réitérable sur un grand nombre de documents et (3) réalisable par un grand nombre d'annotateurs.

4 Systématisation des annotations et modélisation

O. Houdé (1992) relève que la pratique systématisée de l'annotation repose sur un phénomène essentiel de l'activité humaine :

> la catégorisation est la conduite adaptative fondamentale par laquelle les systèmes cognitifs, biologiques ou artificiels découvrent le réel physique et social. Sa fonction normative[10] est la création de classes d'équivalence nécessaires à la transition du continu au discret (manipulable).

C'est bien l'objectif visé par l'annotation des corpus : rendre le texte manipulable, créer des points de vue sur les données, le rôle des métadonnées de contenu étant de permettre l'accessibilité des données pour en faire des objets de la recherche. Dégager et projeter les catégories sur les observables, afin de tester leur pouvoir d'élucidation apparaît comme le fondement de la démarche scientifique, et la pratique de l'annotation s'inscrit donc dans le processus heuristique même de l'activité modélisante.

Pour avancer dans la réflexion sur l'utilité des annotations, il nous faut donc aborder un autre aspect : le fait que les annotations expriment des catégories[11]. Pour mieux

[9] Pour s'en convaincre, on peut penser aux consignes de correction typographique, dont on trouvera un exemple sur http://www.alain.les-hurtig.org/pdf/signes-de-correction.pdf, consulté le 10 septembre 2021.

[10] La *norme* est ici à comprendre comme l'élaboration et l'identification et d'une classe d'objets définie par une norme (par exemple un paradigme descriptif).

[11] L'identification du remplacement d'une minuscule par une majuscule lors de la correction des épreuves d'un texte regroupe une catégorie d'événements qu'on (an)notera systématiquement par le même symbole.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

approcher le phénomène de l'élaboration de catégories et, ensuite, d'un modèle de représentation d'événements, nous repartirons d'exemples très simples, basés sur la correction de copies.

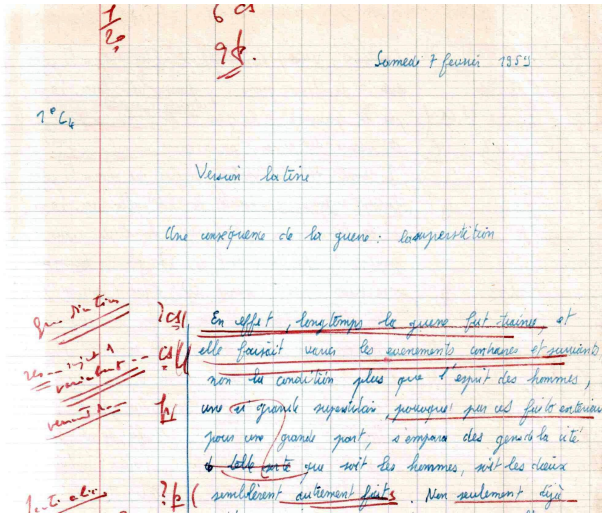


Fig. 2 : Annotations sur une copie de version latine (Archives personnelles)

Le texte initial (ici la production de l'élève) est enrichi par les annotations de l'enseignant. On identifie ici sans difficulté l'empan de texte concerné : il s'agit des zones de texte de l'élève soulignées en rouge. On identifie tout aussi aisément la glose explicative, en rouge, dans la marge gauche de la copie. Enfin, on voit apparaître deux étiquettes, encadrées par nous en vert : contre-sens (cs) et faux-sens (fs), qui ont pour objectif de « rationaliser » la description du type de problème rencontré : on projette des catégories, et, pour ce faire, on a recours à un vocabulaire fixe (fs et cs).

Les étiquettes (métadonnées qu'on retrouvera plus loin sous le nom d'annotations) ont donc pour fonction de *matérialiser un regard construit, catégorisant, sur les données brutes*. Pour que l'annotation d'un document ou d'un ensemble de documents, en tant que processus d'affectation des étiquettes et en tant que résultat de l'opération, soit exploitable, il faut impérativement stabiliser à la fois la *signification* des étiquettes et leurs *conditions d'affectation*.

Là encore, il s'agit d'un phénomène répandu que nous illustrons ici par la présentation d'une typologie de fautes d'orthographe :

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

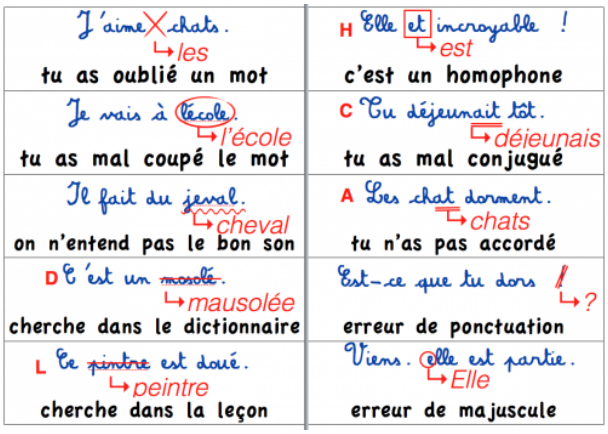


Fig. 3 : Typologie de « fautes d'orthographe » (Archives personnelles)

On voit ici un exemple d'ébauche d'établissement d'un vocabulaire fixe destiné à catégoriser les remédiations (par exemple les différentes façons de mettre en évidence la localisation de l'erreur et le type d'opération à effectuer pour y remédier : D pour « cherche dans le dictionnaire »).

Les exemples donnés jusqu'ici mettent en œuvre différentes pratiques de l'annotation de texte, et permettent d'en vérifier des éléments stables (d'autant plus stables que l'on se rapproche de l'époque actuelle) : (1) localisation (2) choix d'un vocabulaire fixe (3) explicitation de la signification et (4) explicitation des conditions d'attribution des étiquettes.

Ces exemples relèvent tous d'une même activité, celle de la correction du texte[12], dans des perspectives différentes, amenant à des choix d'étiquettes différents ; et c'est là un point crucial : la pratique de l'annotation implique une intention d'exploitation.

Dans le cadre d'une démarche heuristique portant sur des phénomènes langagiers et littéraires (entre autres), les annotations permettent de repérer toutes les occurrences d'une même catégorie d'événements, c'est-à-dire l'identification de paradigmes répondant aux traits définitoires de la catégorie projetée sur le texte par le biais de l'annotation. La pratique la plus connue est celle de l'affectation d'étiquettes morpho-syntaxiques, du type verbe ou déterminant, etc. L'annotation permet aussi d'observer l'insertion des phénomènes annotés en contexte, c'est-à-dire le repérage de leurs conditions d'apparition dans l'ordre syntagmatique, celui du déroulement du discours. L'observation conjointe des deux dimensions, paradigmatique et syntagmatique, peut amener à réviser les définitions des catégories repérées : les affiner, les regrouper, soit, plus généralement, les restructurer.

On entre ainsi dans la fameuse boucle heuristique, démarche expérimentale avec progression en boucle qui permet de tester des hypothèses par l'examen de l'efficacité des modèles construits, de documenter les étapes différentes étapes de la recherche et d'en assurer ainsi la reproductibilité.

[12] Correction au sens large, les annotations visant à la production d'une « meilleure » version du texte.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Reste à examiner l'impact que peut avoir l'environnement numérique sur cette pratique : ce qui relève d'une simple transposition d'un environnement à l'autre, et dans quelle mesure les outils numériques sont *utiles*, tant lors de l'opération d'annotation que lors de l'exploitation de ces annotations, ne serait-ce que pour faire plus vite et mieux la même chose. On peut aussi se demander dans quelle mesure le changement de média peut ouvrir à d'autres perspectives : les environnements numériques ont en effet cette propriété remarquable qu'ils modifient la plasticité des textes, et permettent d'accéder à leur contenu en se dégageant des contraintes d'une lecture linéaire.

5 Les corpus annotés, de l'objet à l'outil de la recherche

La première fonction des pratiques d'annotation est de permettre une recherche d'information ciblée, *filtrée*, qui, dans nos contextes, sera mobilisée lors de l'élaboration des « données de la recherche », ou, le plus souvent dans les faits, pour n'en aborder qu'une étape : l'extraction d'un corpus d'étude. Le parallèle avec les métadonnées documentaires s'impose : « l'objectif de la documentarisation est d'optimiser l'usage du document en permettant un meilleur accès à son contenu et une meilleure mise en contexte » (Salaün, 2007). L'objectif de l'annotation de contenu est similaire : là où la recherche « sur texte brut » ne permet que des repérages très strictement basés sur les chaînes de caractères, la recherche sur textes annotés permet d'accéder au texte par le biais d'annotations qui catégorisent les observables. Ces annotations peuvent être assez variées (Habert, Nazarenko et Salem, 1997), porter sur différentes unités (du caractère au texte – *cf.* TEI), concerner tout ou partie du corpus, être mono- ou multi-couches.

Dans le sens où nous employons le mot ici, les corpus ne sont donc pas seulement des données ou des objets de la recherche, *ils en sont aussi les outils*, dans la mesure où ils sont élaborés de façon à répondre à une intention : c'est la définition de cette intention qui permet, par exemple, de statuer sur la pertinence d'un corpus, ne serait-ce que pour établir dans quelle mesure on peut considérer qu'il est représentatif des phénomènes étudiés. En effet, *les corpus d'étude*[13] *sont des objets fonctionnels, construits à des fins précises.* Soumis au critère de pertinence, et même d'une pertinence « dense » en lien direct avec le projet d'analyse, ils sont construits sur des critères qui optimisent leur propension à manifester quelque chose (un modèle, une matrice) pouvant avoir d'autres manifestations. Par ailleurs, ils peuvent manifester diverses matrices présentant différents points de vue sur un objet, et peuvent provoquer un réaménagement des modèles et de leur expression, les annotations. C'est en cela que le corpus, qui ne saurait être autre chose que construit[14] (et se distingue déjà, dans les faits, des données empiriques réputées brutes ou factuelles[15], même s'il n'est pas annoté), peut jouer son rôle de médiateur entre les faits langagiers et le

[13] Sur la distinction entre corpus de référence et corpus d'études, voir Rastier (2005).

[14] Ce point ne semble plus faire débat depuis le début des années 2000, comme en témoignent ces propos de Mellet (2002) : « le constat s'est partout imposé que le corpus est un objet construit, que ce soit à travers la structuration, l'organisation des données retenues, voire leur enrichissement au moyen des procédures du balisage et de l'étiquetage. Le fait n'est pas nouveau : toute classification, tout tableau synthétique récapitulatif de la philologie traditionnelle offraient aussi cette part d'interprétation implicite des données ».

[15] Il est par ailleurs légitime de douter du fait que les données ne sont pas déjà des objets construits, mais là n'est pas notre propos.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

chercheur, et être ainsi l'objet intermédiaire entre les données empiriques et les modélisations (linguistiques ou autres). *Le corpus ainsi vu n'est pas un donné fixe, c'est un objet paramétré, dynamique et sujet à variation* : il se constitue ainsi en terrain d'expérimentation.

Dans la mesure où elles projettent un point de vue catégorisant sur les données, les annotations contribuent clairement à ce façonnage du corpus en tant qu'objet/outil d'étude. Dans les environnements numériques, cette pratique, sous-jacente à toute opération de modélisation, se trouve « outillée », ce qui a un impact déterminant à la fois sur la masse de données que l'on peut traiter de manière exhaustive, et sur le nombre d'hypothèses modélisantes que l'on peut tester de façon systématique. Dès lors, on peut considérer que le changement d'échelle autorisé par le changement d'environnement a un impact qui ne se limite pas à la dimension quantitative : tester des hypothèses systématiquement et à grande échelle, être en mesure de tester de nombreuses formulations intermédiaires affecte l'activité de recherche de façon qualitative.

6 L'annotation et son exploitation : concepts et outils

Du point de vue conceptuel comme du point de vue pratique, l'outillage de l'activité d'annotation comme de l'exploitation des données annotées repose sur la transposition de procédés anciens, voire antiques. De nouveau, il nous faudra examiner si l'apport de l'environnement numérique se limite à un « saut quantitatif[16] ».

Nous l'avons vu, l'annotation dans les environnements non numériques a une fonction essentielle : dégager des catégories qui sont pertinentes d'un certain point de vue, déterminé par une tâche. En d'autres mots, il s'agit d'identifier des paradigmes de phénomènes afin d'en rationaliser l'étude et le traitement.

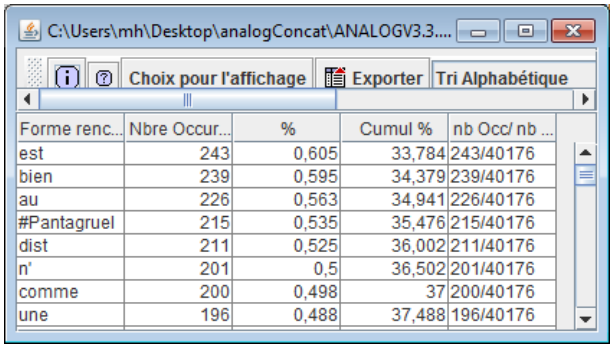
Un autre point essentiel a été souligné : l'identification de ces catégories ne repose pas sur la seule localisation d'un événement, mais sur l'observation de son apparition dans un contexte donné. On l'illustre aisément en morpho-syntaxe : suivant les contextes, « rire » (et le paradigme de ses occurrences) peut-être un nom ou un verbe ; son insertion syntagmatique permet de faire le choix : en tant que nom, il pourra être précédé d'un déterminant et suivi d'un adjectival (*un rire joyeux, un rire qui/que/de …*), en tant que verbe à l'infinitif, son insertion sera tout autre (*voilà qui prête à rire, pas de quoi en rire, ce n'est pas le moment de rire, on peut rire de tout*, etc.) Ce qui est vrai de l'observation des catégories grammaticales l'est de tout autre type de catégorie : l'établissement des définitions de dictionnaire suppose qu'on distingue les emplois en discours, le choix de la bonne traduction d'un mot nécessite de lever les ambiguïtés sémantiques en contexte, etc. Il en va de même des pratiques exégétiques.

[16] Les éléments de réflexion présentés ici sont plus sont exposés plus en détail dans Lay (2019).

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Ce constat n'est pas nouveau et motive des pratiques anciennes, elles aussi largement communes aux pratiques des sciences humaines : l'établissement et l'observation d'index[17] et de concordances[18].

Établir des index et des concordances, tâche souvent incontournable, est généralement assez mécanique *et* extrêmement coûteux en temps, bref, rébarbatif et ingrat. C'est cependant une étape féconde pour la recherche, car elle donne l'occasion de se familiariser intimement avec les données et d'élaborer des hypothèses innovantes. L'environnement numérique permet d'en systématiser la pratique et les exploitations : nous donnerons ici l'illustration du procédé sur une donnée simple et continue, la forme-graphique *Pantagruel* (dans le texte de Rabelais, 1532[19]), puis nous présenterons une généralisation du procédé en l'appliquant aux méta-données et à des séquences d'information.



Forme renc...	Nbre Occur...	%	Cumul %	nb Occ/ nb ...
est	243	0,605	33,784	243/40176
bien	239	0,595	34,379	239/40176
au	226	0,563	34,941	226/40176
#Pantagruel	215	0,535	35,476	215/40176
dist	211	0,525	36,002	211/40176
n'	201	0,5	36,502	201/40176
comme	200	0,498	37	200/40176
une	196	0,488	37,488	196/40176

Fig. 4 : Index des occurrences de *Pantagruel*[20]

[17] Les index sont disponibles depuis l'antiquité, ils peuvent être extrêmement différents les uns des autres, contentons-nous ici d'en évoquer deux, de complexité différente : les index de mots, dits « de bas niveau » car « au ras du texte », et les index de notions, qui supposent qu'on se détache de la littéralité séquentielle, d'une part en prenant de la « hauteur » (un positionnement *méta*, donc), d'autre part en traitant des unités continues ou discontinues (p. ex. un seul mot ou plusieurs mots), qu'elles soient contiguës ou non (p. ex. les différents marqueurs de négation répartis sur une phrase).

[18] La concordance comme instrument de travail textuel est née au XIII^{e} siècle. La première d'entre elles semble être une concordance biblique élaborée par les Dominicains. La tradition de la concordance biblique perdure et est aujourd'hui outillée, voir (http://www.bibliquest.org/concordance-avancee.htm, consultée le 10 septembre 2021). Là où les index permettent de faire apparaître les événements d'une façon très synthétique (nombre d'occurrences d'un item, éventuellement associées à des informations de localisation), la concordance permet d'en présenter toutes les occurrences contextualisées sous une forme « verticalisée », « empilée », propre à outiller l'analyse des contextes (*cf.* infra, la concordance sur *Pantagruel*).

[19] Texte disponible dans la base Epistémon-BVH, http://www.bvh.univ-tours.fr/Epistemon/index.asp, consulté le 10 septembre 2021.

[20] Les index et concordances présentés dans ce document ont été produits à l'aide de Navilog, outil d'annotation manuelle développé à Poitiers par M.-H. Lay. (215 occurrences sur les 40176 mots du texte)

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

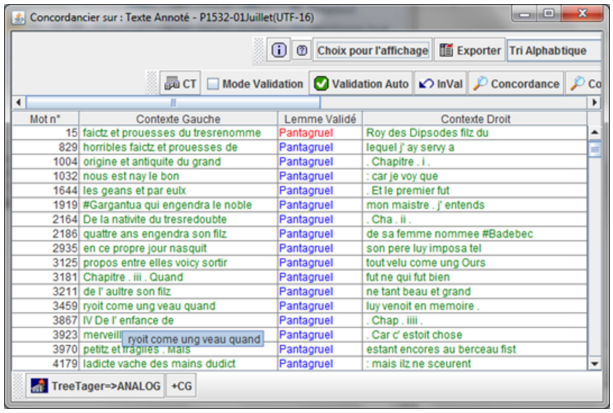


Fig. 5 : Concordance sur les occurrences de la forme pivot *Pantagruel*

Dans leur version la plus simple, index et concordances portent sur les mots-graphiques. Les index en fournissent l'inventaire et les transforment ainsi en *type*, une opération fondamentale de la lexicométrie [voir entre autres Lebart et Salem (1994)]. Il y a, par souci d'économie (interprétative, conceptuelle, mémorielle, …), coïncidence entre le nom du *type* (abstraction qui regroupe toutes les occurrences d'une forme, et par conséquent, première étape d'une vision *méta* des données) et la forme matérielle de sa réalisation : on dit par exemple que « mange » apparaît douze fois, peu importe qu'il soit en italique, en gras, en telle ou telle police de caractère etc. Même dans des contextes simples, la constitution des index suppose de fait une première étape de catégorisation, qui peut être amenée à se généraliser progressivement, déclenchant alors le mécanisme de l'annotation : l'étape suivante sera par exemple de repérer les occurrences de « manger », indépendamment du fait qu'il soit conjugué de telle ou telle façon, ou encore d'annoter le concept de « prise alimentaire », et donc de regrouper toutes sortes d'éléments lexicaux (*manger, ingérer, avaler*, …) ou syntagmatiques (*troubles de la déglutition*) sous une même étiquette.

Dès lors que les textes sont annotés, on peut produire des index et des concordances sur les annotations ou sur des séquences d'événements, faisant alterner les annotations et les formes : de fait, lors des phases d'exploitation, les métadonnées sont des données comme les autres[21].

Pour illustrer ce point, les figures qui suivent (Figures 6 et 7) présentent les concordances et index de la structure [Verbe + *de*] pour l'une, et pour l'autre, différentes façons de présenter les index du concept « DISCOURS » rapporté aux formes et annotations.

[21] On peut par ailleurs se demander si les données de la recherche ne sont pas par définition des métadonnées ; la question est assez aisément perceptible lorsqu'on procède par exemple à la transcription de corpus oraux ou multi-modaux, mais il en va vraisemblablement de même pour les documents textuels : quelle que soit la forme qu'il prend, un « a », manuscrit ou numérique, est la manifestation d'une catégorie, dotée d'attributs circonstanciels (« main », police, effets, etc.). C'est bien le fait qu'il exprime une catégorie (« a » regroupant toutes les occurrences de *a*) qui permet de l'identifier comme unité fonctionnelle constitutive d'un signifiant linguistique.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

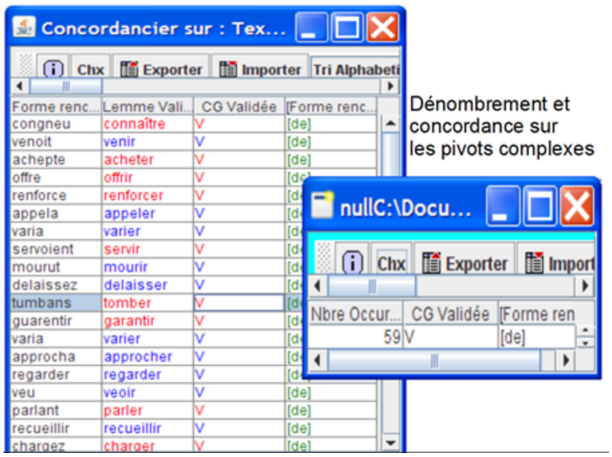


Fig. 6 : Inventaire et dénombrement des formes – prises par la séquence [Verbe+*de*] dans le *Pantagruel* de 1532

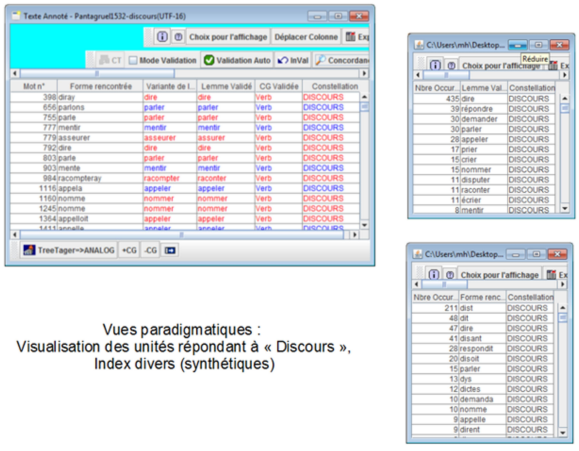


Fig. 7 : Inventaire et dénombrement des formes – prises par les verbes associés à la notion de DISCOURS dans le *Pantagruel* de 1532

Notre propos n'est pas ici de nous attarder plus longuement sur la diversité des extractions et tris possibles, mais de rendre aisément perceptible le fait que les raisons d'utiliser des concordanciers et des outils d'annotation ne font aucun doute : la facilité des repérages et des extractions, la possibilité de faire rapidement de nouveaux tris, sont certes des atouts qui permettent de travailler sur un plus gros volume de données, mais ils permettent surtout de multiplier les hypothèses et de tester plusieurs hypothèses en un temps réduit : un tel changement d'échelle ne se traduit pas en simples termes de volumétrie. Il permet de justifier et de documenter les trajectoires de l'analyse et d'entrer dans le paradigme des sciences expérimentales, élément essentiel lors de la rédaction, par exemple des réponses aux appels à projet.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Les outils proposés reposent sur le fait que l'exploitation de grands corpus (grand signifiant ici « plus important que ce que la mémoire embrasse ») ne saurait s'accommoder d'une lecture linéaire, même si le retour au texte dans son déroulement est toujours nécessaire et généralement implémenté, de façon plus ou moins riche. Il est nécessaire de pouvoir faire apparaître les redondances qui permettent à la fois de définir les phénomènes à annoter (élaborer des catégories), de procéder à des annotations systématiques (projeter des catégories) et de les interpréter, puis de les réaménager, le cas échéant (exploiter les catégorisations). Ainsi, les outils permettent d'assister et de documenter les différentes étapes de la recherche, de la formulation d'hypothèses à leur validation menant (éventuellement) à la validation d'une modélisation :

(1) la formulation d'hypothèses : seule l'observation de toutes les occurrences d'un phénomène (index) en contexte (concordances) permet de formuler une hypothèse sur ce phénomène dans ce contexte (l'observation de tous les contextes permet de voir en quoi l'hypothèse est éventuellement généralisable et justifie la mise en place d'une catégorie) ; disposer de tous les contextes identiques permet de mettre en évidence des propriétés (par exemple distributionnelles) des unités. Cette phase permet donc la conception de catégories, l'élaboration d'un système d'annotation.

(2) La reformulation de l'hypothèse (voire la modification du corpus d'étude) : utilisée en phase d'élaboration du modèle (catégories et relations entre les catégories), la souplesse des outils et leur capacité à traiter très rapidement des masses de données permet aisément la remise en cause du modèle et le réajustement des annotations : la pertinence de cette description sera jugée à l'aune de son utilité déterminée par un point de vue modélisant[22].

(3) La validation/invalidation de l'hypothèse : la détection de redondances rendues préhensibles par l'annotation permet la mise en regard de différents contextes, et permet la formulation d'hypothèses sur le degré de *généralisabilité* de la description retenue. Cette fois, ce ne sont plus les seules occurrences des phénomènes actualisés que l'on peut observer, mais l'occurrence de l'abstraction modélisante. C'est donc l'observation des catégories élaborées qui est rendue possible, tandis que la cohérence et la pertinence du modèle peuvent être interrogées.

[22] Selon les contextes, la modélisation doit pouvoir être modifiée afin d'entrer à nouveau dans le cycle « observation/validation ». Pour le chercheur, ce point est central. Il doit se prémunir du fait que, dès lors qu'un texte est annoté, on risque une forme de circularité, comme le souligne Hunston (2002) : « the categories used to annotate a corpus are typically determined before any corpus analysis is carried out, which in turn tends to limit the kind of question that usually is asked. » L'impossibilité de procéder à ces réajustements peut conduire à s'interroger sur le bien-fondé même de l'étiquetage, sauf à le concevoir comme un procédé dynamique, réitérable tout au long de l'analyse. En effet, l'annotation ne pouvant tenir compte de toutes les dimensions de l'objet, il y a d'un certain point de vue perte d'information. En témoignent les inquiétudes de Sinclair (2005) : « The interspersing of tags in a language text is a perilous activity, because the text thereby loses integrity. » Pourtant, même à imaginer que le corpus *perde en intégrité*, il gagne une chose indispensable à son étude : il *gagne en préhensibilité*, et devient véritablement *champ d'expérimentation*, dès lors que les outils d'exploration (index et concordances) et les outils d'annotation sont construits en étroite relation d'interdépendance.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

7 Conclusion : utilisation raisonnée de l'annotation et indépendance du chercheur

Les observations de Martinet (1979) sur la nécessaire dépendance entre pertinence d'une description et stabilisation d'un point de vue clairement identifié illustrent à quel point il est nécessaire de se poser et de reposer la question de la pertinence de façon systématique, y compris dans les domaines où les consensus peuvent sembler bien établis, comme celui de la description grammaticale. La pratique de l'annotation en environnement numérique (y compris en Traitement Automatique du Langage) pose la question dans les mêmes termes. Habert (2000) note par exemple que « quelle que soit sa richesse, une annotation est cependant toujours orientée par une tâche, même si cela est implicite ». Quant à Leech (2005), il insiste sur ce point : « les annotations sont d'autant plus utiles qu'elles ont été conçues pour une application particulière ».

Les outils d'annotation manuelle ont vocation à assister le chercheur dans cette tâche : affecter, modifier, valider des annotations obtenues d'une façon ou d'une autre. Elles peuvent être le fruit d'une lecture cursive s'accompagnant d'enrichissement progressifs, au fil du texte. Elles peuvent avoir été produites par la simple projection sur le texte d'une ressource de métadonnées (par exemple un thésaurus), ou avoir été produites par un *tagger* automatique. Le point crucial est que l'annotateur doit pouvoir embrasser l'ensemble des occurrences d'une étiquette, afin de valider la cohérence de l'étiquetage. Il lui faut avoir accès à l'ensemble des métadonnées disponibles pour la description d'une unité donnée, et ce afin de valider son choix comme étant le meilleur possible dans un référentiel donné. Il doit aussi pouvoir modifier les métadonnées associées à une unité, soit parce qu'il associe de nouvelles valeurs à une étiquette (par exemple s'il rencontre une nouvelle dénomination pour un personnage de roman qu'il étiquette par ailleurs), soit parce qu'il se rend compte que certaines étiquettes doivent être réorganisées (par exemple introduire une étiquette [type de focale] pourrait amener à un réaménagement d'un thésaurus portant sur la dénomination des lentilles optiques à l'époque de la Renaissance). Ces opérations, de natures assez différentes (elles concernent divers aspects des métadonnées), demandent à être assistées[23] par des briques logicielles dont l'utilisateur souhaite qu'elles soient aisément accessibles depuis son espace de lecture[24]. Il nous semble que le contexte général est aujourd'hui propice au développement de cette nouvelle gamme d'outils, et qu'il est essentiel pour les chercheurs en sciences humaines et sociales de s'emparer de la question pour être associés à leur spécification[25]. Pour que notre usage des corpus entre véritablement dans l'ère numérique, il nous faut œuvrer à mieux intégrer les ingénieurs informaticiens dans nos pratiques de recherches et pas seulement dans les phases de constitution et de stockage des données.

[23] Par exemple, les modifications concernant les catégories et leur définition demandent à être sérieusement accompagnées : dans un système, modifier un élément équivaut à modifier la valeur de tous les autres éléments. Toute modification à ce niveau affecte donc potentiellement la totalité du système.

[24] C'est le parti que nous avons pris pour NaviLog, outil d'annotation manuelle développé à Poitiers en collaboration avec les Bibliothèques Virtuelles du centre de la Renaissance à Tours, et l'ANR textométrie portée par l'ENS-Lyon. (Voir Lay, 2015).

[25] Qu'il s'agisse de la définition des fonctionnalités et modalités d'usage des briques logicielles, ou en termes de « légèreté » de l'appropriation technique permettant l'autonomie du chercheur : installation, formats d'import et d'export, prise en main, etc.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

On peut noter depuis une dizaine d'années l'apparition de l'annotation manuelle comme sujet de publication. Limiter ce phénomène à la montée en puissance de sa pratique ne prend pas en compte un aspect peut-être plus essentiel : il nous semble que l'absence de publications traduisait aussi le fait que l'activité d'annotation manuelle a longtemps été reléguée au statut de tâche ancillaire, du point de vue des « informaticiens » comme du point de vue des « chercheurs en sciences humaines ». Pour présenter les choses de façon caricaturale, il s'agissait, pour les uns comme pour les autres, d'un travail de tâcheron[26], *préalable* au travail scientifique clairement identifiable dans la recherche en algorithmique pour les uns, ou dans les études monographiques pour les autres. Vue ainsi, l'annotation n'a que le statut de simple étape initiale ou intermédiaire, et ne peut prétendre à rien d'autre que d'être assimilée à un « enchaînement d'opérations techniques », quand elle n'est pas purement et simplement disqualifiée. Ainsi, un article, Scott *et al.* (2012) soutiennent la position suivante :

> Annotation studies in Computer Linguistics are generally unscientific : they are mostly not reproductible, make use of too few (and often non-independent) annotators and use guidelines that are often something of a moving target.

Ce point de vue pourrait amener à de longs développements sur un contexte où l'activité scientifique se « technicise » et où son évaluation en tant qu'activité scientifique se limite trop souvent à une lecture technicienne des termes de falsifiabilité des modèles et de reproductibilité des activités : il nous semble pourtant assez clair que l'efficacité technique n'est pas à assimiler à une preuve scientifique de la robustesse des modèles[27].

Quant à l'aspect mouvant de l'affectation des catégories ou même de leur définition, nous espérons avoir montré qu'elles ne sont pas la manifestation d'un manque de rigueur scientifique, mais le fondement même d'une activité d'élucidation du réel, car ce qui sous-tend l'activité modélisante, c'est la reconfiguration des représentations jusqu'à l'identification d'une solution satisfaisante. Traditionnellement, l'annotation manuelle en permet l'expression, tandis que l'environnement numérique en permet la montée en puissance et en modifie la portée. Les outils de calcul permettent des sauts quantitatifs qui, alliés à la plasticité de l'information dite dématérialisée, amènent un saut qualitatif et provoquent un changement de paradigme dont les SHS peuvent s'emparer, tant pour affiner les concepts et les conditions de leur mise en œuvre, que pour contribuer à l'élaboration des outils et à l'établissement de protocoles d'usages, de guides de bonnes pratiques. Nous avons exposé en quoi, méthodologiquement et techniquement, l'annotation manuelle nous semble prendre une place centrale dans les SHS et contribue à leur visibilité institutionnelle en tant que « science ». Les années qui viennent diront la place que les chercheurs souhaitent lui réserver dans leurs pratiques.

[26] Il en va de même de la transcription des corpus ou la rétro-conversion de documents.

[27] L'efficacité reproductible de l'orientation des panneaux solaires en fonction du mouvement du soleil dans le ciel ne prouve pas que le soleil se déplace.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352