Étienne Cuvelier, Sébastien de Valeriola, Céline Engelbeen

Identification automatique des sources des notices zoologiques du *Speculum naturale* de Vincent de Beauvais

Volume 1, nº 1 (2020), p. 19-42.

<http://roia.centre-mersenne.org/item?id=ROIA_2020__1_1_19_0>

Revue Ouverte d'Intelligence Artificielle
Volume 1, n° 1, 2020, 19-42

Identification automatique des sources des notices zoologiques du *Speculum naturale* de Vincent de Beauvais

Étienne Cuvelier[a], Sébastien de Valeriola[a], Céline Engelbeen[a]

[a] ICHEC - Brussels Management School, Laboratoire Quaresmi, Boulevard Brand Whitlock, 2, 1150 Bruxelles.
E-mail : etienne.cuvelier@ichec.be, sebastien.devaleriola@ichec.be, celine.engelbeen@ichec.be.

Résumé. — Avec son encyclopédie intitulée *Speculum maius*, le dominicain du XIIIe siècle Vincent de Beauvais tente de constituer une synthèse générale du savoir. Pour ce faire, il rassemble des renseignements provenant d'une multitude de sources différentes, chrétiennes et païennes, antiques et médiévales. La plupart des notices de son œuvre contiennent une mention explicite des sources dont elles sont inspirées, à la différence de beaucoup des encyclopédies médiévales. Cette caractéristique permet d'utiliser le *Speculum maius* comme base d'expérimentation, et de lui appliquer des techniques d'apprentissage supervisé et de fouille de textes dans le but de relier automatiquement les notices encyclopédiques à leurs sources. Dans cet article, nous nous livrons à cet exercice pour les livres zoologiques de cette encyclopédie et nous analysons ensuite les apports, les limites et les perspectives des résultats obtenus dans l'optique d'une application future à d'autres encyclopédies dont les notices ne mentionnent pas leurs sources.

Mots-clés. — Fouille de textes, Apprentissage supervisé, Encyclopédie médiévale.

1. Introduction

D'après certains historiens (par exemple [20]), les milieux intellectuels de l'Occident du XIIe siècle sont le théâtre d'une véritable « renaissance ». Il est en tout cas indéniable que cette période marque une profonde mutation au niveau culturel. Celle-ci se poursuit au siècle suivant, considéré parfois comme « l'âge d'or » des encyclopédies médiévales (voir [22], nuancé par exemple dans [14]).

Avec l'idée d'organiser l'ensemble des connaissances de l'époque (des textes antiques mais aussi des traductions de textes grecs ou arabes vers le latin, qui arrivent en Occident à cette époque) certains auteurs regroupent et compilent l'ensemble des savoirs sous la forme d'encyclopédies (pour un panorama de l'encyclopédisme médiéval, voir [3]).

Une question naturelle se pose dès lors, celle de l'identification des sources utilisées par les encyclopédistes médiévaux. Ces encyclopédies formant un corpus très volumineux, cette démarche est longue et fastidieuse, raison pour laquelle elle n'a pas été entreprise par l'historiographie de manière complète et globale. C'est ce qui a motivé le développement de notre méthodologie d'identification automatique basée sur des outils de fouille de textes (*text mining*) que nous présentons dans cet article. Nous ne saurions trop insister sur le fait que cet outil ne remplace pas l'expertise de l'historien, mais accompagne celui-ci dans la procédure d'identification, notamment en lui proposant une liste ordonnée de sources plausibles.

1.1. Le choix du corpus de Vincent de Beauvais

Parmi les encyclopédistes médiévaux les plus influents, une triade d'auteurs contemporains se dégage : le dominicain Thomas de Cantimpré, le franciscain Barthélémy l'Anglais et le dominicain Vincent de Beauvais. Décédé en 1264, ce dernier rédige le *Speculum maius*, qui forme l'une des encyclopédies les plus importantes du Moyen Âge [43, p. 210].

Nous avons choisi de travailler sur celui-ci, et ce pour plusieurs raisons. Premièrement, le texte du *Speculum*, dans une édition moderne (celle de Douai, datant de 1624), est disponible en version électronique sur le site SourcEncyMe [42]. L'acquisition du corpus n'a donc pas posé de problème.

Deuxièmement, la forme des notices qui le composent est particulièrement bien adaptée à l'exercice que nous effectuons. En effet, d'une part celles-ci apparaissent comme de relativement courts ensembles de phrases bien séparés les uns des autres, que l'ordinateur peut considérer comme des entités indépendantes. D'autre part, Vincent place en tête de chacune de ces « rubriques » ce que l'historiographie appelle un « marqueur » [44], qui renseigne le lecteur sur la source utilisée. Dans plusieurs manuscrits conservés, ce marqueur est par ailleurs mis en évidence par rapport au reste du texte, par exemple par l'utilisation d'une couleur particulière, comme dans l'exemple donné à la figure 1.1.

La présence de ce marqueur possède un intérêt particulier dans le cadre qui est le nôtre, car il nous renseigne *a priori* sur la source effectivement utilisée par Vincent (et par conséquent sur la qualité des identifications obtenues automatiquement par l'ordinateur), et nous place donc dans un cadre similaire à celui de l'apprentissage supervisé. Néanmoins, ces marqueurs ne sont pas complets (Vincent renseigne le plus souvent sa source en ne donnant que l'auteur, rarement le numéro du livre correspondant), et semblent parfois inexacts (comme nous le verrons *infra*). Les différentes techniques que nous utilisons, ainsi que les méthodes existantes dont nous nous inspirons dans cette étude, seront discutées *infra*.

Les marqueurs découpent ainsi chaque chapitre du *Speculum* en sections. Dans la suite de cet article, les notices de Vincent seront désignées par un triple $V(l; c; s)$, où l est le numéro du livre, c le numéro du chapitre et s le numéro de la section, tels qu'ils

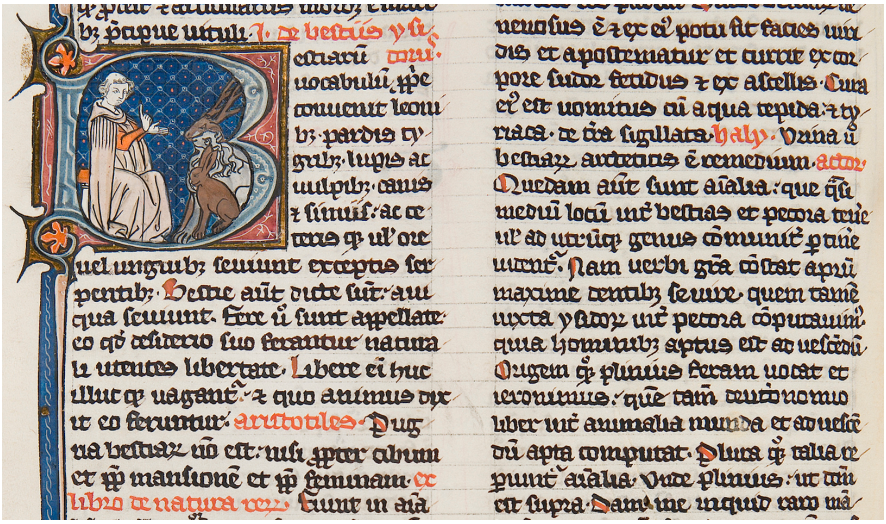

FIGURE 1.1. Marqueurs médiévaux (en rouge) dans un manuscrit du *Speculum naturale* (Tournai, Séminaire épiscopal, ca 1280, f° 120v, livre 18). Photographie © KIK-IRPA, Bruxelles.

apparaissent dans le découpage effectué sur SourcEncycMe (qui suit le découpage originel et donc celui de l'édition de Douai, à l'exception de l'ajout des sections).

Troisièmement, la façon dont Vincent rédige son encyclopédie rend l'identification automatique possible (contrairement à la situation d'autres auteurs, comme Albert le Grand, voir [18, p. 244]). Notre auteur réutilise en effet les extraits dont il tire ses informations sans leur apporter de modifications majeures. Bien souvent, il réemploie tels quels des blocs de texte presque *verbatim*, comme l'illustre l'exemple du tableau 1.1.

V(17; 4; 2)	Pline [27, p. 116, l. 19-23]
Pisces branchias habentes non existimant quidam anhelitum reddere ac per vices recipere, ut Aristoteles	[...] Nec piscium branchias habentes anhelitum reddere ac per uices recipere existimant [...], in qua sententia fuisse Aristotelem uideo et multis persuasisse doctrina insignibus.

TABLE 1.1. Comparaison d'une notice de Vincent et de sa source.

Il s'ouvre d'ailleurs sur ce procédé dans le prologue de l'ouvrage [32, p. 149-156] :

> « *Il y a tant de livres, une telle multitude, le temps de la vie est si bref et la mémoire si faible, que l'esprit humain ne peut s'approprier tout ce qui a été écrit. Pour ces raisons,* [...] *j'ai choisi, selon mes possibilités, des extraits de presque tous les livres que j'ai lus, œuvres*

> *des docteurs chrétiens et auteurs païens, poètes et philosophes* [...]
> *je ne procède pas à la manière d'un docteur, ni d'un auteur de traité,*
> *mais toujours comme un compilateur* [...]. »

Il faut cependant remarquer que, comme nous le verrons *infra*, le texte de Vincent ne correspond souvent pas exactement au texte de sa source (en tout cas dans les versions qui nous sont connues). Les écarts consistent tantôt en des différences minimes (ordre des mots, constructions grammaticales, etc.), tantôt en des insertions ou des remaniements plus importants. Ces derniers sont néanmoins plus rares.

Le *Speculum maius* comporte 80 livres répartis en trois parties [3, p. 1262] : le *Speculum naturale* et le *Speculum historiale*, présents dès la première version (qui date de 1244-1246), et le *Speculum doctrinale*, qui apparaît avec la deuxième version (1250). Nous avons choisi d'élaborer notre outil à partir de la première d'entre elles, et plus précisément des livres de celle-ci qui traitent de zoologie (soit les livres 16 à 22, dont le contenu est donné dans le tableau 1.2, d'après [44]).

N°	Sujet	Nombre de chapitres
16	Oiseaux	171
17	Monde marin	146
18	Animaux domestiques	98
19	Animaux sauvages	139
20	Reptiles et insectes	179
21	Anatomie des animaux	66
22	Activités et génération des animaux	68

TABLE 1.2. Sujets des livres du *Speculum naturale* que nous considérons.

La raison de ce choix tient en ce que les sources de cette partie du *Speculum naturale* sont relativement bien connues par l'historiographie (voir [44]).

1.2. Les sources considérées

Comme nous l'avons déjà signalé *supra*, les sources utilisées par Vincent pour écrire les livres zoologiques du *Speculum naturale* sont bien connues par l'historiographie, en tout cas d'un point de vue macroscopique (c'est-à-dire qu'on sait assez bien à quelles œuvres il fait référence, sans pourtant généralement savoir quels passages sont utilisés). Le tableau 1.3 donne la distribution des références mentionnées dans les marqueurs médiévaux.

Trois des sources qui y apparaissent appellent un commentaire. La première, signalée « ? » dans le tableau, correspond aux 146 notices qui ne portent pas de marqueur. La deuxième, signalée « Auctor », correspond aux notices dans lesquelles Vincent a inséré des observations qu'il a effectuées lui-même ou des renseignements qu'il tient de ses contemporains (à propos de ce phénomène, voir [31, 24]).

Auteur	Total	Auteur	Total	Auteur	Total
Plinius	979	Pythagoras	10	Rufus	2
Liber Nat. R.	405	Radulphus	10	Strabus	2
Aristoteles	376	Andromachus	7	Val. Maximus	2
Isidorus	309	**Comestor**	8	Zeno	2
?	146	Helinandus	7	Sacra Historia	1
Auctor	182	Glossa	6	Aemilius	1
Ambrosius	113	Gregorius	6	Plautus	1
Avicenna	208	Vegetius	6	Liber empyricus	1
Solinus	91	Hieronymus	5	Cicero	1
Physiologus	86	Papias	5	Suetonius	1
Glossa [Bible]	64	Physicus	5	Guill. de Conchis	1
Dioscorides	60	Achildes	4	Basilius	1
Palladius	56	Belbetus	4	Serapion	1
Isaac	47	Platearius	4	Bali.	1
Hali	44	Seneca	4	Rabanus	1
Iorath	40	Virgilius	4	Iuvenalis	1
Aesculapius	38	Aviarium	3	Polybius	1
Razi	17	Ovidius	3	Algazel	1
Philosophus	13	Albertus	2	**Orosius**	1
Alexander	19	Fulgentius	2	Origenes	1
Constantinus	15	Horatius	2	Greg. Nissenus	1
Lucanus	15	Hesychius	2	Democritus	1
Augustinus	10	Martialis	2		

TABLE 1.3. Liste des sources potentielles ; en gras, les sources utilisées dans ce projet.

La troisième, signalée par « Liber Nat. R. », devrait à première vue correspondre à l'encyclopédie de Thomas de Cantimpré, qui circule alors sans que le nom de son auteur ne lui soit attaché [44, p. 150]. L'historiographie a néanmoins remarqué que certaines de ces notices sont inspirées d'une autre encyclopédie, intitulée *Liber de naturis rerum*, rédigée entre 1220 et 1240, dont l'auteur ne nous est pas connu (cette encyclopédie est généralement désignée comme celle de Pseudo-John Folsham, du nom d'un carmélite anglais du XIV^e^ siècle à laquelle elle a été erronément attribuée, [1, p. i-lxii, surtout p. xxxv et suivantes]). L'une des hypothèses avancées pour expliquer cette apparente confusion de Vincent entre les deux ouvrages est que celui-ci aurait travaillé d'après un manuscrit concernant les deux textes copiés l'un à la suite de l'autre [30, p. 31-33]. Nous avons donc utilisé les deux textes dans notre analyse, en les concaténant tout simplement. Cette source « double » sera désignée par l'anagramme commun aux deux titres, LDNR.

Il est bien entendu nécessaire, pour mettre en œuvre la comparaison désirée, de disposer des textes des sources que Vincent a potentiellement consultées. Malheureusement, ceux-ci ne sont pas tous disponibles en édition électronique. Il nous a donc

fallu sélectionner, au sein de la longue liste des auteurs de référence donnés par les marqueurs de Vincent, ceux qui étaient facilement exploitables. Ceux-ci sont identifiés dans le tableau 1.3 par des caractères gras. Nous obtenons ainsi un total de 13 524 notices-sources. Les éditions de référence sont données en annexe de cet article.

Puisque nous avons travaillé dans une optique similaire à celle de l'apprentissage supervisé, et que nous disposons pas de l'ensemble des sources mentionnées par Vincent, il nous a fallu restreindre notre corpus aux notices dont le marqueur médiéval indiquait un auteur dont nous disposions des œuvres. Le total des notices de Vincent se porte alors à 2 411.

2. Méthodologie

La détection automatique de la source potentielle d'une notice du *Speculum naturale* est de l'ordre de la ré-utilisation de textes (*Text re-use*). Celle-ci est aussi bien utilisée en histoire ([8], [23], [9]), dans la détection de plagiat ([26], [36], [33]) ou le journalisme ([11]).

Les techniques utilisées peuvent être basées sur les n-grammes ([25], [6]) ou sur les recouvrements de sous-chaînes de caractères ([48]). Une autre approche se base sur l'alignement de textes dans différentes langues, en se basant sur les longueurs de ceux-ci ([7], [17]) ou en utilisant les *cognates* ([41]) qui sont des termes similaires dans les langues des textes à comparer.

Notons que dans notre cas, les techniques basées sur les comparaisons de chaînes de caractères ne peuvent fonctionner car comme on le constate dans l'exemple du tableau 1.1, entre une source et son utilisation par Vincent de Beauvais, les formes des mots peuvent varier. L'approche par n-grammes s'appuyant sur l'ordre des mots n'est pas complètement adéquate car cet ordre n'est pas encore absolu dans le latin médiéval. Enfin l'alignement de texte suppose un relatif parallélisme entre un texte et sa traduction ce qui n'est pas toujours le cas avec le *Speculum naturale*, Vincent de Beauvais reformulant par moment certains morceaux des phrases de la source. De par les variations illustrées dans le tableau 1.1, une étape de la lemmatisation s'avère absolument nécessaire. Cette étape est souvent un précédent à l'approche par sac de mots ([35]), qui utilise une matrice documents-termes pour construire une représentation vectorielle des documents, la comparaison entre ces derniers se faisant alors souvent en utilisant directement la similarité cosinus sur lignes de la matrice. Si cette dernière approche peut convenir aux textes écrits dans une langue où l'ordre des mots n'est pas complètement contraint, ce n'est pas notre cas, le latin médiévale étant dans une situation intermédiaire et évolutive.

C'est pour ces diverses raisons que nous proposons une approche différente, basée sur des matrices de tailles raisonnables et tenant relativement compte de l'ordre des lemmes. L'idée de notre méthodologie d'identification automatique est que chacune des comparaisons notice-source potentielles produit une métrique de comparaison, et qu'ensuite, sur base de l'ensemble des résultats, on sélectionne la source potentielle la plus probable (c'est-à-dire celle dont la métrique est la plus élevée). La source

détectée automatiquement de la sorte doit ensuite être comparée à celle reprise dans le marqueur médiéval. Dans le cas où ces deux sources correspondent (attention, comme dit dans l'introduction, Vincent ne renseignant en général pas les livres et les chapitres mais uniquement les auteurs de ses sources, nous ne pouvons que vérifier que nous avons identifié le même auteur), l'identification est interprétée comme une réussite. Dans les autres cas, elle est comptée comme une erreur. L'efficacité de la méthodologie proposée peut ensuite être calculée simplement comme le pourcentage de réussites.

Cette section méthodologique est essentiellement divisée en deux parties : une partie explicitant les pré-traitements appliqués aux textes avant toute analyse et ensuite une partie explorant les différentes métriques de comparaison testées.

2.1. Pré-traitement

Comme souvent en fouille de textes les techniques d'analyse ne sont pas appliquées directement aux textes originaux mais à des versions transformées de ceux-ci. L'idée générale de ces transformations est d'apurer les textes de ce qui peut rendre plus difficile les opérations de comparaison. La première opération est de distinguer les différents éléments du texte considéré en repérant les différents séparateurs que sont les espaces et les signes de ponctuation, en anglais on parle de *tokenisation*. Cette première étape permet de passer de l'état de données non structurée (un texte) à l'état de données structurées (un vecteur de mots). Une fois les différents éléments du texte isolés il est possible d'éliminer les éléments de structuration du texte, à savoir, la ponctuation et les mots de liaisons. Ne reste alors du texte que ses éléments signifiants. Les traitements que nous effectuons sur le matériel originel ont pour but de faciliter les comparaisons entre une notice de l'encyclopédie considérée et les différentes sources potentielles. Or entre le texte de la source et sa présence dans le *Speculum naturale*, Vincent de Beauvais a, vraisemblablement comme c'est souvent le cas lorsque l'on synthétise du texte, dû reformuler les éléments de connaissances qu'il intégrait. Ce qui a comme conséquence immédiate que la forme des termes à comparer entre une notice et ses sources potentielles peuvent être différentes : par exemple, les verbes peuvent être conjugués différemment et les mots accordés aussi de façon différente. Ces différences de formes pourraient compliquer les comparaisons informatisées puisque, de façon très basique, la succession des lettres a changé. C'est pour solutionner ce problème que la fouille de textes recourt à la lemmatisation : cette opération permet de passer de la forme fléchie (accordée ou conjuguée) du lemme à sa forme non conjuguée et non accordée. De façon imagée cela revient à trouver la forme du mot ou verbe telle qu'on la trouve dans un dictionnaire.

La figure 2.1 illustre le processus de pré-traitement en utilisant le package `R stringr` [19] pour la partie tokenisation et le logiciel `Treetagger` [38] pour la partie lemmatisation. Remarquons dans la même figure que `Treetagger` présente ses résultats en trois colonnes : la première colonne est celle des mots originaux, la deuxième colonne donne le rôle du mot et la troisième colonne donne son lemme.

Dans notre cas les portions de texte qui sont considérées sont les sections du *Speculum naturale*. Notons qu'on ne peut pas passer sous la barre de la granularité

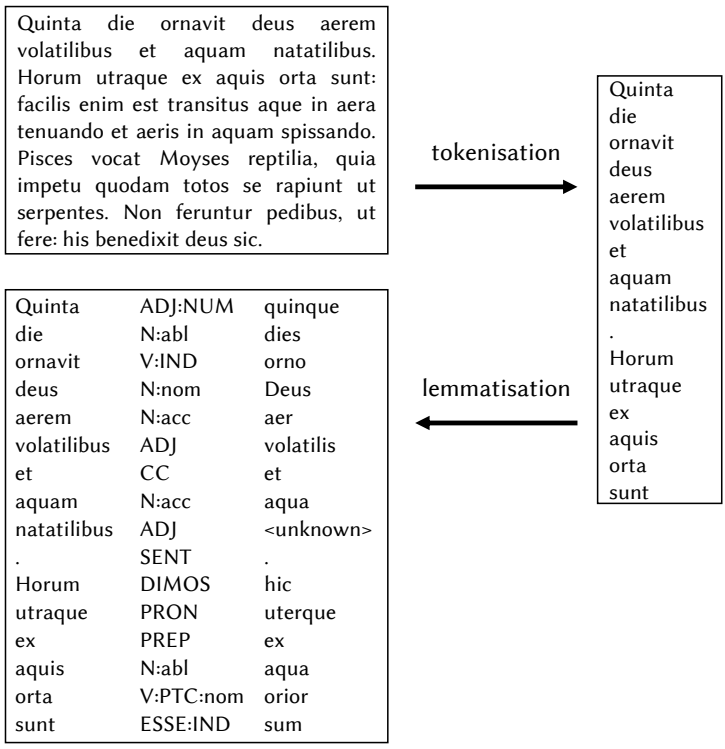

FIGURE 2.1. Processus de tokenisation/lemmatisation.

des sections indiquées par Vincent de Beauvais lui-même, parce que le niveau du dessous est la phrase et que l'encyclopédiste ne respecte pas forcément les phrases de ses sources.

2.2. ANALYSES

Une fois que nous sommes en possession des versions vectorisées (tokenisées) et lemmatisées des notices de Vincent de Beauvais et des sources potentielles sélectionnées, il est alors nécessaire de déterminer une métrique de comparaison qui permette de retrouver la source réelle avec la plus grande probabilité.

2.2.1. *Signatures de distances de mots*

Nous utiliserons un petit exemple pour illustrer notre démarche. On trouvera dans le tableau 2.1 une notice extraite de Vincent notée $V(17; 9; 1)$ en accord avec la notation définie précédemment (page 22), une notice d'Isidore [2, p. 493, l. 19-20] et une notice de Pline [27, p. 132, l. 3-4] référencées selon la même convention que celle de Vincent

$V(17;9;1)$	Pisces dicti sunt a pascendo.
$I(12;6;1)$	Pisces dicti unde et pecus, a pascendo scilicet.
$P(9;1;41)$	Est paruus admodum piscis adsuetus petris,

TABLE 2.1. Les trois notices servant d'exemple.

(pour plus de clarté) : $I(12;6;1)$ et $P(9;1;41)$. Ces deux dernières constituent deux exemples de sources potentielles avec lesquelles sont comparées les notices de Vincent.

$V(17;9;1)$	piscis dictus sum pasco
$I(12;6;1)$	piscis dictus unde pecus pasco scilicet
$P(9;1;41)$	sum paruus admodum piscis assuetus petra

TABLE 2.2. Les notices lemmatisées et nettoyées.

Les tableaux 2.2 et 2.3 regroupent les mêmes notices, lemmatisées et nettoyées de la ponctuation et des mots de liaison dans le premier cas et sous forme vectorisée dans le second cas. Par forme vectorisée nous entendons que pour chaque notice nous considérons le vecteur (au sens de tableau à une dimension) des lemmes la composant, avec une position déterminée par leur ordre d'apparition dans la notice originale. Notons que, dans le cas où un lemme apparaît plusieurs fois dans une notice, c'est la position de sa première apparition dans l'ordre de la lecture qui est utilisée (comme le formalise la définition 2.4). Accessoirement dans ce dernier tableau, et ce afin de faciliter la compréhension, nous avons coloré chaque mot de la notice $V(17;9;1)$ dont nous cherchons la source d'une couleur qui lui est propre, cette même couleur étant utilisée dans les autres notices $I(12;6;1)$ et $P(9;1;41)$.

	1	2	3	4	5	6
$V(17;9;1)$	piscis	dictus	sum	pasco		
$I(12;6;1)$	piscis	dictus	pecus	pasco	scilicet	
$P(9;1;41)$	sum	paruus	admodum	piscis	assuetus	petra

TABLE 2.3. Les notices vectorisées.

Sur base de ces versions vectorisées des notices nous pouvons définir aisément la position d'un lemme dans une notice.

DÉFINITION 2.1. — *Nous définissons la fonction $pos(l,n)$ qui donne la position du lemme l dans la notice n :*

$$pos(l,n) = i$$

signifiant donc que le lemme l se situe à la position i dans la notice n, sachant que la numérotation commence à 1 et se termine à k si le nombre de mots dans n est égale à k, i.e. : $1 \leqslant i \leqslant k$. Si le lemme l ne se trouve pas dans la notice n, alors $pos(l,n)$ n'est pas définie.

Ainsi par exemple, en utilisant le tableau 2.3 on peut facilement calculer que

$$pos(\text{pasco}, V(17; 9; 1)) = 4 \text{ et } pos(\text{pasco}, I(12; 6; 1)) = 4 \tag{2.1}$$

alors que $pos(\text{pasco}, P(9; 1; 41))$ n'est pas définie.

Sur base de la position d'un lemme dans une notice nous pouvons aisément calculer la distance qui sépare deux lemmes présents dans une notice, et cette distance nous sert de mesure de dissimilarité entre les deux lemmes considérés.

Définition 2.2. — *Une mesure de dissimilarité dans un ensemble V entre deux éléments $u, v \in V$ est toute fonction $d : V \times V \to \mathbb{R}^+$ qui satisfait les deux propriétés suivantes :*

Séparation : $d(u, u) = 0$,
Symétrie : $d(u, v) = d(v, u)$.

Définition 2.3. — *La distance entre deux lemmes l_1 et l_2 présents dans la notice n est donnée par*

$$d(l_1, l_2, n) = |pos(l_1, n) - pos(l_2, n)|. \tag{2.2}$$

Si au moins l'un des deux lemmes n'est pas présent dans la notice n alors la distance $d(l_1, l_2, n)$ n'est pas définie.

Notons que cette définition implique que la distance d est transparente aux décalages constants des lemmes d'une notice. Ainsi, si on crée une nouvelle notice n' en insérant k lemmes au début d'une notice n, les distances entre les lemmes de n' qui étaient déjà présents dans n n'est pas modifiée : pour tous i, j, on a

$$d(l_{i+k}, l_{j+k}, n') = d(l_i, l_j, n).$$

Cette propriété n'est évidemment vérifiée que si les lemmes ajoutés à n' sont tous différents des lemmes de n.

Sur base de cette notion de dissimilarité nous proposons de constituer des signatures qui nous permettront d'encoder à la fois la présence ou l'absence d'un lemme dans une notice, mais aussi sa position relative dans celle-ci. Pour cela nous commençons par définir la matrice des distances des lemmes d'une notice de référence, qui dans cet article est systématiquement une notice du *Speculum naturale*.

Définition 2.4. — *Soit une notice de référence r contenant k lemmes différents. La matrice de distances des lemmes de r dans la notice n est la matrice $k \times k$, notée $D_r(n)$, et dont les éléments sont donnés par*

$$[D_r(n)]_{i,j} = d(l_1, l_2, n), \tag{2.3}$$

où :

- *l_1 et l_2 sont deux lemmes présents dans n,*
- *$i = pos(l_1, r)$,*
- *$j = pos(l_2, r)$,*

les indices étant déterminés par rapport à l'ordre des lemmes dans la notice de référence r. Si un mot se retrouve plusieurs fois dans une notice, c'est sa première apparition dans l'ordre de la lecture qui est prise en compte.

On trouve dans le tableau 2.4 les matrices de distances des lemmes de $V(17;9;1)$ dans les notices $I(12;6;1)$ et $P(9;1;41)$, mais aussi dans la notice $V(17;9;1)$ elle-même qui, servant de notice de référence, verra sa matrice de distances être aussi une matrice de référence.

Nous obtenons donc des matrices telles que :

- l'absence d'un lemme l d'une notice de référence r dans une notice n étant signifiée par l'absence de nombres sur la ligne et la colonne correspondante ($\text{pos}(l,n) = i$) dans la matrice correspondante $D_r(n) : \forall j \in \{1,\dots,k\} : [D_r(n)]_{i,j} \notin \mathbb{R}$ et $[D_r(n)]_{j,i} \notin \mathbb{R}$. Dans le cas contraire il existe au moins un indice de colonne contenant un nombre : $\exists j \in \{1,\dots,k\} : [D_r(n)]_{i,j} = [D_r(n)]_{j,i} \in \mathbb{R}$;
- pour les lemmes de r présents dans n, la matrice $D_r(n)$ contient toutes leurs distances respectives. Ces différentes distances dépendant de l'ordre des lemmes dans la notice n, cette matrice de distance permet donc d'encoder cet ordre. Ainsi la simple permutation de deux lemmes modifie les lignes et les colonnes correspondant aux dits lemmes dans la matrice $D_r(n)$.

	piscis	dictus	sum	pasco
piscis	0	**1**	**2**	**3**
dictus	1	0	**1**	**2**
sum	2	1	0	**1**
pasco	3	2	1	0

	piscis	dictus	sum	pasco
piscis	0	**1**	**-**	**3**
dictus	1	0	**-**	**2**
sum	-	-	-	**-**
pasco	3	2	-	0

	piscis	dictus	sum	pasco
piscis	0	**-**	**3**	**-**
dictus	-	-	**-**	**-**
sum	3	-	0	**-**
pasco	-	-	-	-

TABLE 2.4. Matrices des distances des lemmes des notices reprises dans le tableau 2.3 : (a) $D_{V(17;9;1)}(V(17;9;1))$, (b) $D_{V(17;9;1)}(I(12;6;1))$, (c) $D_{V(17;9;1)}(P(9;1;41))$.

Comme dans toute matrice de dissimilarité, seules $k \times (k-1)/2$ informations sont pertinentes puisque la matrice est symétrique et sa diagonale est toujours nulle. Compte tenu de cette propriété, nous avons décidé d'extraire l'information pertinente se trouvant dans la partie triangulaire supérieure et de sérialiser celle-ci sous forme de vecteurs de signature.

DÉFINITION 2.5. — *Soient une notice de référence r contenant k lemmes différents et $D_r(n)$ la matrice des distances des lemmes de r dans une notice n, alors la signature de distances de n par rapport à r est le vecteur constitué des éléments de la partie triangulaire supérieure de $D_r(n)$.*

Cette signature de distances est notée $\vec{D}_r(n)$ et ses éléments constitutifs sont les $d_{i,j} = [D_r(n)]_{i,j}$ (où $j > i$) disposés dans l'ordre lexicographique des indices.

$$\vec{D}_{V(17;9;1)}(V(17;9;1)) = (1;2;3;1;2;1)$$
$$\vec{D}_{V(17;9;1)}(I(12;6;1)) = (1;-;3;-;2;-)$$
$$\vec{D}_{V(17;9;1)}(I(9;1;41)) = (-;3;-;-;-;-)$$

TABLE 2.5. Les signatures de distances des notices.

Même si ces signatures permettent d'avoir sous forme numérique les informations concernant la présence ou l'absence d'un lemme et l'ordre relatif des lemmes de référence dans une notice de comparaison, les lemmes pour lesquels ces distances ne sont pas définies ne rendent pas cette signature complètement satisfaisante. C'est pourquoi nous allons transformer ces dissimilarités en similarités.

2.2.2. *Signatures de similarités de mots*

DÉFINITION 2.6. — *Une mesure de similarité entre deux éléments de $u, v \in V$ est toute fonction $s : V \times V \to \mathbb{R}^+$ qui satisfait les propriétés suivantes :*

Séparation : $s(u, u) = k$, *où k est une constante,*
Symétrie : $s(u, v) = s(v, u)$,
Maximalité : $s(u, v) \leqslant s(u, u) = k$.

Toute mesure de dissimilarité d peut être transformée en mesure de similarité s, et vice et versa, au travers d'une fonction ϕ strictement décroissante :

$$s(u, v) = \phi(d(u, v)) \text{ et } d(u, v) = \phi^{-1}(s(u, v)) \tag{2.4}$$

avec comme condition que $\phi(0) = k$. La fonction de densité gaussienne est un exemple de fonction ϕ utilisée dans ce cas et c'est cette dernière que nous utiliserons.

DÉFINITION 2.7. — *Soit une notice de référence r contenant k lemmes différents et $\vec{D}_r(n)$ la signature de distances de n par rapport à r, alors sa signature de similarités de fenêtre $w \in \mathbb{R}_0^+$, notée $\vec{S}_{r,w}(n)$, est le vecteur de $\mathbb{R}^{k\times(k-1)/2}$ constitué des éléments suivants :*

$$\left[\vec{S}_{r,w}(n)\right]_i = \begin{cases} e^{-\frac{1}{2}\left(\frac{\left[\vec{D}_r(n)\right]_i}{w}\right)^2} & \text{si } \left[\vec{D}_r(n)\right]_i \in \mathbb{R}, \\ 0 & \text{sinon.} \end{cases} \tag{2.5}$$

Cette mesure de similarité a l'avantage d'être à valeurs dans l'intervalle $[0, 1]$: 0 correspond au cas d'une dissimilarité complète et 1 au cas d'une similarité complète. Évidemment le paramètre de fenêtrage w est un élément important à calibrer puisque, au vu des propriétés de la courbe gaussienne, il détermine la « vitesse » de décroissance des similarités calculées à partir de dissimilarités croissantes.

Ce paramètre w nous aide à créer des signatures qui permettent de distinguer, parmi les six situations possibles, les trois cas de figure dans lesquelles une notice r du *Speculum naturale* peut se trouver par rapport à sa source n :

(1) ***n* est source de *r*** et la notice r est une copie exacte, soit d'une notice source n, soit d'une partie de ladite source ;
(2) ***n* est source de *r*** mais la notice r est une version légèrement reformulée du contenu d'une notice source n, soit d'une partie de ladite source ;
(3) ***n* est source de *r*** mais la notice r est le résultat de la collation de plusieurs sources n_i, et donc il existe *a priori* plusieurs sous-parties de r se trouvant dans une des situation décrites ci-dessus ;
(4) ***n* n'est pas source de *r*** et les lemmes de r qui se retrouvent dans n le sont dans des positions relatives assez différentes de leurs positions dans r ;
(5) ***n* n'est pas source de *r*** mais les lemmes de r qui se retrouvent dans n le sont dans des positions relatives assez similaires à leurs positions dans r ;
(6) ***n* n'est pas source de *r*** et aucun lemme de r ne se retrouve dans n.

La situation 6 est évidemment facile à repérer : si aucun lemme de r ne se trouve dans n, alors aucun des $\left[\vec{D}_r(n)\right]_i$ n'est défini, et par conséquent le vecteur $\vec{S}_{r,w}(n)$ est le vecteur nul.

Le cas précédent est lui aussi facile à repérer si la notice n n'a que peu de lemmes de r, puisque dans ce cas le vecteur $\vec{S}_{r,w}(n)$ a d'autant plus de valeurs nulles qui le différencient de $\vec{S}_{r,w}(r)$. Les cas des notices n n'étant réellement pas sources de r mais ayant beaucoup de lemmes de r dans des positions relatives proches de celles de r sont évidemment des cas très difficiles à distinguer par une technique automatique. Un choix adéquat du paramètre w doit nous aider à distinguer le quatrième cas de figure des deux premiers (le troisième cas pouvant être vu comme une version « dégénérée » des cas 1 et 2), puisque dans cette situation la fonction définie en (2.5) donnerait des valeurs plus faibles pour les $\left[\vec{S}_{r,w}(n)\right]_i$, les $\left[\vec{D}_r(n)\right]_i$ étant plus grands. D'une part, une valeur de w faible facilite la discrimination des cas. D'autre part, une valeur élevée implique un risque de confusion entre le cas 2 et le cas 4.

Plusieurs choix ont été testés pour w :

- $w = k$, où k est le nombre de lemmes uniques dans la notice de référence r,
- $w = l$, où l est le nombre de lemmes uniques dans la source potentielle n,
- $w = \min(k, l)$.

C'est le choix $w = l$, qui, expérimentalement, s'est révélé être la valeur permettant d'obtenir le meilleur taux d'identification de la source.

On trouve dans les tableaux 2.6 et 2.7 les signatures de similarités calculées à partir des signatures de distances du tableau 2.5 avec w respectivement égale à 5 et 6 les nombres de lemmes différents dans les notices de comparaison $I(12; 6; 1)$ et $P(9; 1; 41)$.

$\vec{v_i} = \vec{S}_{V(17;9;1),5}(V(17;9;1)) = (0{,}98; 0{,}92; 0{,}83; 0{,}98; 0{,}92; 0{,}98)$
$\vec{i} = \vec{S}_{V(17;9;1),5}(I(12;6;1)) = (0{,}98; 0; 0{,}84; 0; 0{,}92; 0)$

TABLE 2.6. Les signatures des similarités normalisées des notices.

$\vec{v_p} = \vec{S}_{V(17;9;1),6}(V(17;9;1)) = (0{,}99; 0{,}95; 0{,}88; 0{,}99; 0{,}95; 0{,}99)$
$\vec{p} = \vec{S}_{V(17;9;1),6}(P(9;1;41)) = (0; 0{,}88; 0; 0; 0; 0)$

TABLE 2.7. Les signatures des similarités normalisées des notices.

À cette étape il s'agit, dans notre exemple, de distinguer automatiquement quelle signature parmi $\vec{i}$ et $\vec{p}$ produit la similarité la plus élevée avec celle de $V(17;9;1)$. Une mesure de similarité assez naturelle pour les vecteurs, déjà évoquée dans la section 2, est la similarité *cosinus* qui calcule le cosinus de l'angle entre deux vecteurs. On en trouve en (2.6) la version pour nos vecteurs de signatures de similarités.

$$\text{sim}(r,n) = \frac{\vec{S}_{r,w}(r) \cdot \vec{S}_{r,w}(n)}{||\vec{S}_{r,w}(r)||.||\vec{S}_{r,w}(n)||} \tag{2.6}$$

Plus les valeurs de deux signatures sont différentes, plus leur similarité *cosinus* est faible. L'idée est que la source véritable doit être la notice ayant la valeur $\text{sim}(r,n)$ la plus importante.

On trouve le détails des calculs pour notre exemple dans les expressions (2.7) à (2.14), ce qui permet au néophyte en calcul vectoriel de suivre pas à pas la démarche. Nous commençons par calculer les numérateurs de l'expression (2.6), c'est-à-dire les produits scalaires de $\vec{v_i}$ et de $\vec{i}$ (2.7) ainsi que de $\vec{v_p}$ et de $\vec{p}$ (2.8).

$$\vec{v_i} \cdot \vec{i} = 0{,}98 \cdot 0{,}98 + 0{,}92 \cdot 0 + 0{,}83 \cdot 0{,}84 + 0{,}98 \cdot 0 + 0{,}92 \cdot 0{,}92 + 0{,}98 \cdot 0 = 2{,}51 \tag{2.7}$$

$$\vec{v_p} \cdot \vec{p} = 0{,}99 \cdot 0 + 0{,}95 \cdot 0{,}88 + 0{,}88 \cdot 0 + 0{,}99 \cdot 0 + 0{,}95 \cdot 0 + 0{,}99 \cdot 0 = 0{,}83 \tag{2.8}$$

Ensuite nous calculons les normes des différents vecteurs (2.9-2.12).

$$||\vec{v_i}|| = \sqrt{0{,}98^2 + 0{,}92^2 + 0{,}83^2 + 0{,}98^2 + 0{,}92^2 + 0{,}98^2} = \sqrt{5{,}28} = 2{,}29 \tag{2.9}$$

$$||\vec{i}|| = \sqrt{0{,}98^2 + 0 + 0{,}84^2 + 0 + 0{,}92^2 + 0} = \sqrt{2{,}51} = 1{,}58 \tag{2.10}$$

$$||\vec{v_p}|| = \sqrt{0{,}99^2 + 0{,}95^2 + 0{,}88^2 + 0{,}99^2 + 0{,}95^2 + 0{,}99^2} = \sqrt{5{,}49} = 2{,}34 \tag{2.11}$$

$$||\vec{p}|| = \sqrt{0 + 0{,}88^2 + 0 + 0 + 0 + 0} = 0{,}88 \tag{2.12}$$

Enfin nous calculons les deux similarités (2.13-2.14).

$$\text{sim}(V(17;9;1), I(12;6;1)) = \frac{2{,}51}{2{,}29 \cdot 1{,}58} = 0{,}69 \tag{2.13}$$

$$\text{sim}(V(17;9;1), P(9;1;41)) = \frac{0{,}83}{2{,}34 \cdot 0{,}88} = 0{,}40 \tag{2.14}$$

Pour une notice r et un ensemble $\{n_1, \ldots, n_m\}$ de sources potentielles nous proposons comme source la plus probable celle donnée par le résultat suivant :

$$\text{Source}(r) = \arg \max_{n_1,\ldots,n_m} (\text{sim}(r, n_1), \ldots, \text{sim}(r, n_m)), \tag{2.15}$$

où la fonction arg max renvoie comme valeur l'argument (parmi les valeurs $n_1, ..., n_m$) qui donne la valeur maximum parmi les valeurs $\text{sim}(r, n_1), \ldots, \text{sim}(r, n_m)$. Donc pour notre exemple si nous notons $\mathcal{V} = V(17; 9; 1)$, $\mathcal{I} = I(12; 6; 1)$ et $\mathcal{P} = P(9; 1; 41)$, alors comme illustré en (2.16) nous en déduisons que la source de la notice du *Speculum naturale* de notre exemple provient de la notice d'Isidore considérée.

$$\text{Source}(\mathcal{V}) = \arg\max(\text{sim}(\mathcal{V}, \mathcal{I}), \text{sim}(\mathcal{V}, \mathcal{P})) = \mathcal{I} = I(12; 6; 1) \tag{2.16}$$

Bien entendu, dans notre exemple nous n'avons considéré que deux sources potentielles pour des raisons de lisibilité mais dans le processus général c'est bien 13 524 notices candidates qui sont considérées. On trouve dans l'algorithme 1 la marche à suivre générale. Sur base des marqueurs médiévaux nous pouvons calculer un taux global d'identifications correctes en comparant pour chaque notice de référence r la source potentielle s et ledit marqueur.

Algorithme 1 : Algorithme de recherche de la source potentielle d'une notice r du *Speculum naturale*. La fonction nlemmes(n) donne le nombre de lemmes uniques d'une notice n.

Données : une notice de référence r, un ensemble de sources potentielles $\{n_1, \cdots, n_m\}$.
Résultat : s source potentielle de r.

Lemmatiser et nettoyer r et $\{n_1, \cdots, n_m\}$.
Calculer $D_r(r)$ et $\vec{D}_r(r)$
pour $i \leftarrow 1$ **à** m **faire**
    w = nlemmes(n_i)
    Calculer $\vec{S}_{r,w}(r)$
    Calculer $D_r(n_i)$, $\vec{D}_r(n_i)$ et $\vec{S}_{r,w}(n_i)$
    Calculer $\text{sim}(r, n_i)$
fin
$s = \text{Source}(r) = \arg\max_{n_1,\cdots,n_m}(\text{sim}(r, n_1), \cdots, \text{sim}(r, n_m))$

Cette recherche des sources des notices reste malgré tout un peu intensive d'un point de vue numérique puisqu'il s'agit de comparer chacune des notices du *Speculum naturale* (soit 2 411 notices) à chacune des sources potentielles (soit 13 524 sources candidates). Ce qui explique pourquoi, si la plupart des traitements ont été réalisés sous `R` [34], les parties les plus intensives l'ont été via une implémentation en `C++` grâce au package `Rcpp` [16].

3. Résultats

Nous pouvons maintenant présenter les résultats que nous avons obtenus à l'aide des méthodes décrites dans la section précédente. L'utilisation de la similarité *cosinus* permet d'atteindre un taux global d'identifications correctes de 87,34 %.

Soulignons au passage que la méthodologie permet de débrouiller la confusion qui règne parmi les marqueurs *Liber de natura rerum* (430 notices au total) et que nous avons mentionnée *supra*. Ces notices peuvent en effet être attribuées de façon précise soit au *Liber de natura rerum* de Thomas de Cantimpré (393 notices, soit 91 %), soit au *Liber de naturis rerum* du Pseudo-John Folsham (37 notices, soit 9 %). Parmi ces dernières, citons par exemple la notice $V(20; 83; 1)$ de Vincent, qui traite de deux espèces d'abeilles. La similarité *cosinus* est particulièrement élevée pour la notice correspondante du *Liber de naturis rerum* (0, 934, [1, p. 274, l. 1557 ; p. 285, l. 1594]), l'attribuant ainsi de manière claire au Pseudo-John Folsham. Il est intéressant de remarquer que la source qui obtient la deuxième similarité *cosinus* (0, 9045) est une notice d'Aristote [47, p. 132, l. 35 ; p. 141, l. 25], qui est sans doute l'inspiration du Pseudo-John Folsham.

Il est nécessaire d'analyser le taux global d'identifications correctes et de tenter de comprendre pourquoi notre méthodologie a associé à certaines notices des sources différentes de celles qui sont indiquées par les marqueurs médiévaux. Le reste de cette section est dévolu à cet exercice. Deux types de phénomènes expliquent les identifications apparemment erronées : ceux qui sont liés à la nature du corpus étudié et ceux qui sont liés à notre approche. Nous passerons ceux-ci en revue successivement.

3.1. Phénomènes liés à la nature du corpus

Le taux global d'identifications correctes que nous venons de mentionner ne rend pas compte de la distribution inégale des résultats en fonction de plusieurs caractéristiques des notices. Parmi celles-ci, on trouve l'attribution à un auteur-source effectuée par les marqueurs médiévaux. Comme le montre le graphe de la figure 3.1, tous les auteurs-sources ne sont en effet pas logés à la même enseigne. Par exemple, seules 66,09 % des 115 notices provenant d'après Vincent de l'œuvre d'Ambroise ont effectivement été attribuées à celui-ci par l'ordinateur, alors que ce taux est égal à 92,25 % pour les 942 notices de Pline.

Pour comprendre ce phénomène, nous nous sommes posé la question des auteurs indiqués par la méthodologie *cosinus* en cas d'incohérence avec le marqueur médiéval : quels auteurs « attirent » les notices incorrectement attribuées, ou, en d'autres termes, quelles notices sont difficilement distinguées par la similarité *cosinus* ? Le tableau 3.1 donne la distribution des attributions automatiques (colonnes) en fonction des auteurs indiqués par Vincent (lignes).

Par exemple, on lit sur la première ligne que, parmi les notices dont le marqueur indique Ambroise, 76 ont été automatiquement attribuées à celui-ci, 2 à Aristote, 1 à Palladius, 1 à Pline et 36 au LDNR.

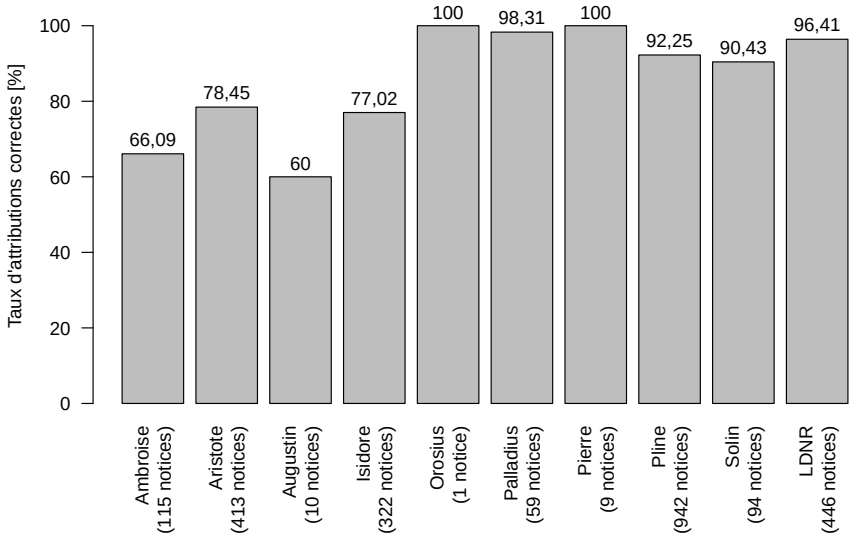

FIGURE 3.1. Ventilation du taux d'identifications correctes en fonction des auteurs indiqués par les marqueurs médiévaux.

	Ambroise	Aristote	Augustin	Isidore	Orosius	Palladius	Pierre	Pline	Solin	LDNR
Ambroise	76	2				1		1		36
Aristote	2	324						3		84
Augustin		1	6							3
Isidore	4	9	1	248	1		1	2	4	52
Orosius					1					
Palladius						58		1		
Pierre							9			
Pline		9		5		5	1	870	2	49
Solin	1	1						1	85	6
LDNR		13	1		1			1		430

TABLE 3.1. Distribution des attributions automatiques (colonnes) en fonction des auteurs indiqués par Vincent (lignes).

La source qui pose le plus de problème est clairement le LDNR. Cette observation n'est pas très surprenante : les deux ouvrages qui sont désignés sous ce nom sont eux-mêmes des encyclopédies, dont les auteurs (Thomas de Cantimpré et un anonyme)

utilisent pour rédiger leurs notices des sources dont beaucoup sont communes avec celles qu'emploie Vincent (Thomas de Cantimpré liste lui-même ses sources dans le prologue de son encyclopédie, voir [15, 62]). Il est donc naturel que certaines notices du LDNR ressemblent fort à celles de Vincent, comme le montre l'exemple donné dans le tableau 3.2.

$V(20; 125; 2)$	Thomas [4, p. 302, l. 7-10] cos = 0, 7434	Ambroise [37, p. 195, l. 2-6] cos = 0, 7432
In exiguo cicadarum gutture dulcis est cantilena, quarum cantibus estu medio rumpunt arbusta, eo quod magis canent meridianis caloribus; quo priorem aerem spiritu attrahunt, eo cantus clariores reddunt.	[...] et hoc quidem dulcis est in exiguo cycadis gutture cantilena. Quarum cantibus, ut Ambrosius dicit, in meridiano estu arbusta rumpuntur, eo quod magis canore meridianis caloribus; quo puriorem aerem per id temporis attrahit spiritu, eo cantus resonant clariores. [...]	Quam dulcis etiam in exiguo cicadis gutture cantilena, quarum cantibus medio aestu arbusta rumpuntur, eo quod magis canorae meridianis caloribus, quo puriorem aerem id temporis attrahunt spiritu, eo cantus resonant clariores. [...]

TABLE 3.2. Comparaison d'une notice de Vincent (dont le marqueur médiéval est « Ambrosius ») avec les sources ayant obtenu les deux meilleures similarités *cosinus*.

Cette explication se trouve confirmée lorsqu'on considère les notices qui sont incorrectement attribuées au LDNR (c'est-à-dire celles qui sont comptabilisées dans la dernière colonne du tableau 3.1) et qu'on regarde de plus près les sources qui ont obtenu la deuxième meilleure similarité *cosinus* (comme donné dans le tableau 3.2). Celles-ci correspondent en effet aux marqueurs médiévaux dans 144 cas sur 230 (63 %). Ces identifications « presque correctes » seraient donc tout à fait correctes si nous ne les mettions pas sur le même pied que les notices d'encyclopédistes contemporains de Vincent de Beauvais.

On peut néanmoins se demander pourquoi la similarité entre deux copies d'une même source est plus élevée que la similarité entre l'une de ces copies et la source elle-même. La raison tient sans doute au moins en partie à une question d'état des textes qui sont aujourd'hui disponibles et que nous avons par conséquent utilisés dans notre étude. Ainsi, dans le cas de l'exemple donné dans le tableau 3.2, il est probable que Vincent et Thomas aient utilisé pour rédiger leurs notices respectives des versions du texte d'Ambroise qui ne correspondent pas tout à fait à celle qui nous sert de référence. Plus de huit siècles séparent en effet les deux encyclopédistes du père de l'église, durant lesquels l'œuvre de celui-ci a pu subir des modifications (au cours de copies successives, par exemple) qui ont été corrigées dans la version que nous connaissons aujourd'hui, mais pas dans celle qui est connue au XIII^{e} siècle.

L'incohérence entre l'information donnée par le marqueur médiéval et l'attribution automatique peut dans certains cas être imputée à Vincent lui-même. Il arrive en effet que le marqueur médiéval donne des renseignements imparfaits.

Ceux-ci sont parfois imprécis, comme par exemple dans le cas de certaines notices qui semblent inspirées chacune de plusieurs sources différentes, mais dont le marqueur n'indique qu'une source unique. Par exemple, la notice $V(17; 9; 1)$ commence par l'indication « Isidorus », alors que seule la première phrase (qui donne l'étymologie du terme *piscis*) provient effectivement des *Etymologiae* d'Isidore de Séville, le reste pouvant être rapproché d'un passage du LDNR de Thomas [4, p. 102-108]. Cette incomplétude pose question, d'autant plus que Vincent indique dans certains cas que sa notice est inspirée de plusieurs sources différentes (par exemple $V(16; 56; 3)$, dont le marqueur est « Plinius ubi supra et etiam ex naturis rerum »). Elle amène parfois à d'apparentes absurdités : la notice $V(17; 29; 5)$ est attribuée à Pline par Vincent (« Plinius libro IXo »), alors que celle-ci intègre un ichtyonyme arabisé (*alphoraz*, pour poisson-écume ; voir à ce sujet [10, 61]), une impossibilité pour l'œuvre de Pline qui a été transmise aux érudits de l'Occident Médiéval directement en latin (voir par exemple [40]). La deuxième partie de cette notice semble en réalité provenir encore une fois du LDNR de Thomas [4, p. 253-254].

Les renseignements donnés par Vincent sont aussi parfois (partiellement) erronés. Ainsi, certaines notices dont l'attribution ne correspond pas au marqueur médiéval ressemblent très fort à la source qui leur a été attribuée automatiquement (elles possèdent d'ailleurs une similarité *cosinus* élevée). Un examen plus approfondi de ces cas-là suggère que c'est le marqueur médiéval qui pose problème. Par exemple, la notice $V(21; 46; 3)$, qui donne une propriété physiologique du cerf, est clairement inspirée de Pline [27, p. 254, l. 28-29 ; p. 255, l.1], tandis que Vincent renseigne Aristote au début de sa notice. L'*Historia animalium* arrive en troisième position en terme de similarité *cosinus* (0, 3649), très loin derrière Pline (0, 816) et le LDNR (0, 5164). De deux choses l'une : soit l'encyclopédiste s'est trompé en rédigeant le marqueur, soit la version du texte dont nous disposons pour la traduction de l'ouvrage d'Aristote ne correspond pas à celle dont disposait Vincent.

3.2. Phénomènes liés à notre approche

D'autres identifications incorrectes peuvent être expliquées par les limites de notre méthodologie. Mentionnons d'abord le découpage des sources en unités textuelles. Afin d'identifier le passage précis d'une œuvre utilisé par Vincent pour rédiger une notice, nous utilisons des subdivisions (chapitres, sections et paragraphes) qui sont parfois le fait des éditeurs modernes. Il arrive donc que le texte auquel le marqueur fait référence soit réparti sur deux subdivisions de la source. C'est par exemple le cas de $V(17; 7; 3)$, où Vincent indique « Plinius libro Xo », mais pour laquelle l'identification automatique ne donne pas de très bons résultats parce que le passage en question est constitué de deux parties différentes dans notre base de données (qui correspondent à [27, p. 200, l. 5-19] et [27, p. 200, l. 20-33]).

Le découpage du *Speculum naturale* lui-même, imposé par la donnée des marqueurs médiévaux, pose aussi des problèmes lorsqu'il aboutit à des notices de longueur inhabituelle. Certaines sont très courtes (par exemple $V(17; 114; 4)$, composée de quatre mots seulement) : le nombre de mots parait alors insuffisant pour que la similarité soit très parlante. D'autres sont très longues (par exemple $V(16; 164; 1)$, dans laquelle Vincent décrit les propriétés médicinales des œufs de différentes races d'oiseaux, qui compte plus de 700 mots), et obtiennent des similarités trop élevées avec certaines sources qui ne leur sont en réalité pas apparentées.

Les difficultés que rencontre notre méthodologie face aux longues notices expliquent sans doute l'hétérogénéité du taux d'identifications correctes par rapport aux livres du *Speculum naturale* que nous considérons. Comme le montre la première moitié du tableau 3.3, les résultats sont bien meilleurs pour les livres 16-20 pour que les livres 21-22 (la différence entre les taux d'identifications correctes est significative avec une p-valeur de $0,0035$).

Groupe	A	B
Livres	16-20	21-22
Nombre de notices	2203	208
Taux d'identifications correctes	89,03 %	84,02 %
Nombre de mots (moyenne)	79	193
Nombre de mots (médiane)	52	165

TABLE 3.3. Comparaison des taux d'identifications correctes pour les différents livres du *Speculum naturale* que nous considérons, et statistiques des nombres de mots (les moyennes et médianes sont calculées sur toutes les notices composant chaque groupe de livres).

Or, comme le montre la deuxième moitié du même tableau, les notices des livres du groupe A sont bien plus courtes que celles des livres de groupe B. Cette différence tient sans doute à la nature des livres en question : la plupart des notices des livres du groupe A consistent en des descriptions de races particulières d'animaux, tandis que les notices des livres du groupe B contiennent des discours plus généralistes sur des sujets transversaux, comme la reproduction des animaux (voir le tableau 1.2).

4. Conclusion

Au terme de cette balade dans les pages zoologiques de Vincent de Beauvais, il est temps de revenir sur quelques aspects de la méthodologie que nous avons implémentée. Nous pensons que les apports de celle-ci peuvent être regroupés sous deux aspects. Il s'agit d'une part du point de vue technique, c'est-à-dire de la perspective *digital humanities*. Notre étude démontre la faisabilité de l'identification automatique des sources des notices encyclopédiques. Les résultats que nous avons obtenus peuvent

certainement être améliorés à plusieurs niveaux. Certains éléments à perfectionner posent des questions techniques qui ne nous semblent pas triviales. Comment par exemple nous affranchir de la limitation imposée par la subdivision des sources en entités textuelles ? D'autres découpages comme ceux évoqués dans [33] peuvent être envisagés.

D'autre part, notre méthodologie est susceptible d'enrichir le champ historiographique de l'encyclopédisme médiéval. Dans le cas que nous avons traité, il s'est agi d'identifier les sources des notices de Vincent de Beauvais, ou plutôt de vérifier l'exactitude des informations fournies par l'auteur lui-même. Les quelques conclusions historiques auxquelles nous sommes arrivés (dans la section 3) – qui sont anecdotiques – n'exploitent que très partiellement les identifications effectuées par l'ordinateur. Nous pensons que l'examen minutieux par un historien des listes de sources proposées automatiquement pourrait mener à d'autres enseignements quant à la façon dont le dominicain manie les textes qu'il considère comme des références.

Nous prévoyons par ailleurs d'appliquer ces techniques à d'autres parties de l'œuvre de Vincent de Beauvais, en choisissant un ou plusieurs ensembles de livres traitant de sujets connexes (par exemple les arbres et plantes avec les livres 9 à 14, les pierres avec le livre 8, etc.). Mais la plus intéressante extension de nos résultats réside certainement dans l'application de la méthodologie à un ensemble de notices sans marqueurs, afin de découvrir cette fois les sources de notices encyclopédiques *a priori* indéterminées (c'est en un sens ce que nous avons fait avec les marqueurs LDNR, comme indiqué au début de la section 3). Par exemple, nous pourrions tenter d'identifier les sources utilisées par Barthélémy l'Anglais dans son *Liber de proprietatibus rerum*, dont les notices ne portent pas d'indication systématique, contrairement à celles que nous avons traitées ici.

En 1990, Onno Boonstra écrivait « The historian who refuses to use a computer as being unnecessary, ignores vast areas of historical research and will not be taken serious anymore » [5]. Si on peut douter de l'universalité de cette assertion, il est indéniable que l'utilisation de méthodes automatisées ouvre de nouvelles perspectives en histoire. C'est dans une optique semblable que nous aimerions inscrire ce projet et ses suites.

Annexe : éditions employées

Les œuvres suivantes ont été utilisées : l'*Hexaemeron* d'Ambroise de Milan [37], l'*Historia animalium* d'Aristote dans sa traduction latine par Michel Scot [45, 46, 47], les *Confessiones*, le *De civitate Dei* et le *De trinitate* d'Augustin d'Hippone [21, 12, 13], le livre 12 des *Etymologiae* d'Isidore de Séville [2], les *Historiarum adversum paganos libri VII* d'Orosius [49], l'*Opus agriculturae* de Palladius [39], l'*Historia scholastica* de Pierre le Mangeur [28], l'*Historia naturalis* de Pline l'Ancien [27], le *Collectanea rerum memorabilium* de Solin [29], le *Liber de natura rerum* de Thomas de Cantimpré [4] et le *Liber de naturis rerum* du Pseudo-John Folsham [1].

Bibliographie

[1] D. Abramov, « "Liber de naturis rerum" von Pseudo-John Folsham - Eine moralisierende lateinische Enzyklopädie aus dem 13. Jahrhundert », phdthesis, Universität Hamburg, 2003, http://ediss.sub.uni-hamburg.de/volltexte/2011/5030/.

[2] J. Andr (éd.), *Isidorus Hispalensis, Etymologiae XII*, Les Belles Lettres, Paris, 1986.

[3] B. Beyer de Ryke, « Le miroir du monde : un parcours dans l'encyclopédisme médiéval », *Revue belge de philologie et d'histoire* **81** (2003), nº 4, p. 23-40.

[4] H. Boese (éd.), *Thomas Cantimpratensis, Liber de natura rerum*, De Gruyter, Berlin et New York, 1973.

[5] O. Boonstra, L. Breure & P. Doorn, *Historische Informatiekunde*, Verloren, Hilversum, 1990.

[6] A. Z. Broder, « Identifying and Filtering Near-Duplicate Documents », in *Annual Symposium on Combinatorial Pattern Matching* (Berlin, Heidelberg), Springer Berlin Heidelberg, 2000, p. 1-10.

[7] P. F. Brown, J. C. Lai & R. L. Mercer, « Aligning sentences in parallel corpora », in *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 1991, p. 169-176.

[8] M. Büchler, G. Crane, M. Moritz & A. Babeu, « Increasing recall for text re-use in historical documents to support research in the humanities », in *International Conference on Theory and Practice of Digital Libraries*, Springer, 2012, p. 95-100.

[9] M. Büchler, A. Gessner, T. Eckart & G. Heyer, « Unsupervised detection and visualisation of textual reuse on ancient Greek texts », *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* **1** (2010), nº 2, p. 1-17.

[10] G. Clesse, « Thomas de Cantimpré et l'Orient. Les sources arabes dans les chapitres zoologiques du Liber de natura rerum », *Reinardus. Yearbook of the International Reynard Society* **25** (2013), p. 53-77.

[11] P. Clough, R. Gaizauskas, S. S. L. Piao & Y. Wilks, « Meter : Measuring text reuse », in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2002, p. 152-159.

[12] « De civitate Dei [édition non identifiée] », http://www.thelatinlibrary.com/august.html.

[13] « De trinitate [édition non identifiée] », http://www.thelatinlibrary.com/august.html.

[14] I. Draelants, « La question ou le débat scolastique comme formes du discours scientifique dans les encyclopédies naturelles du XIIIe siècle : Thomas de Cantimpré et Vincent de Beauvais », *Scientiarum Historia : Tijdschrift voor de Geschiedenis van de Wetenschappen en de Geneeskunde* **31** (2005), nº 1, p. 125-153.

[15] ———, « La science naturelle et ses sources chez Barthélémy l'Anglais et les encyclopédistes contemporains », in *Bartholomeus Anglicus, De proprietatibus rerum. (...) Lateinischer Text und volkssprachige Rezeption* (B. Van den Abeele & H. Meyer, éds.), Brepols, Turnhout, 2006, p. 43-99.

[16] D. Eddelbuettel & R. François, « Rcpp : Seamless R and C++ Integration », *Journal of Statistical Software* **40** (2011), nº 8, p. 1-18.

[17] W. A. Gale & K. W. Church, « A program for aligning sentences in bilingual corpora », *Computational linguistics* **19** (1993), nº 1, p. 75-102.

[18] M. J. Gerhardt, « Zoologie médiévale. Préoccupations et procédés », in *Methoden in Wissenschaft und Kunst des Mittelalters* (A. Zimmermann & R. Hoffmann, éds.), Miscellanea Medievalia, nº 7, De Gruyter, Berlin, 1973.

[19] W. Hadley, « stringr : Simple, Consistent Wrappers for Common String Operations », 2017, R package version 1.2.0, https://CRAN.R-project.org/package=stringr.

[20] C. H. Haskins, *The Renaissance of the Twelfth Century*, Harvard University Press, Cambridge, 1927.

[21] P. Knöll (éd.), *Confessionum libri XIII*, Corpus Scriptorum Ecclesiasticorum Latinorum, nº 33, Tempsky et Freytag, Vienne et Leipzig, 1896.

[22] J. Le Goff, « Pourquoi le XIIIe siècle a-t-il été plus particulièrement un siècle d'encyclopédisme ? », in *L'enciclopedismo medievale* (M. Picone, éd.), Longo, Ravenna, 1994, p. 23-40.

[23] J. Lee, « A computational model of text reuse in ancient literary texts », in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, p. 472-479.

[24] S. Lusignan & M. Paulmier-Foucart, « Vincent de Beauvais et l'histoire du *Speculum maius* », *Journal des Savants* **1-2** (1990), nº 1, p. 97-124.

[25] C. Lyon, J. Malcolm & B. Dickerson, « Detecting short passages of similar text in large document collections », in *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 2001.
[26] H. A. Maurer, F. Kappe & B. Zaka, « Plagiarism-a survey », *J. UCS* **12** (2006), n° 8, p. 1050-1084.
[27] C. Mayhoff (éd.), *Plini Secundi Naturalis historiae libri XXXVII*, Teubner, Leipzig, 1875.
[28] J.-P. Migne (éd.), *Petrus Comestor, Historia scholastica*, Migne, Paris, 1855, col. 1049-1722 pages.
[29] T. Mommsen (éd.), *Caii Julii Solini Collectanea rerum memorabilium*, Weidmann, Berlin, 1895.
[30] S. Moureau, « Les sources alchimiques de Vincent de Beauvais », *Spicae, Cahiers de l'Atelier Vincent de Beauvais* **2** (2012), p. 5-118.
[31] M. Paulmier-Foucart, « L'actor et les auctores : Vincent de Beauvais et l'écriture du *Speculum majus* », in *Auctor et auctoritas : invention et conformisme dans l'écriture médiévale. Actes du colloque tenu à l'Université de Versailles-Saint-Quentin-en-Yvelines, 14-16 juin 1999* (M. Zimmermann, éd.), Mémoires et documents, n° 59, École des chartes, Paris, 2001, p. 145-160.
[32] M. Paulmier-Foucart & M.-C. Duchenne, *Vincent de Beauvais et le Grand miroir du monde*, Brepols, Turnhout, 2004.
[33] M. Potthast, M. Hagen, T. Gollub, M. Tippmann, J. Kiesel, P. Rosso, E. Stamatatos & B. Stein, « Overview of the 5th international competition on plagiarism detection », in *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, CELCT, 2013, p. 301-331.
[34] R Core Team, « R : A Language and Environment for Statistical Computing », 2017, R Foundation for Statistical Computing, Vienna, Austria `https://www.R-project.org/`.
[35] G. Salton & M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., New York, 1986.
[36] M. A. Sanchez-Perez, G. Sidorov & A. F. Gelbukh, « A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014 », in *CLEF (Working Notes)*, Citeseer, 2014, p. 1004-1011.
[37] K. Schenkl (éd.), *Sancti Ambrosii opera*, Corpus Scriptorum Ecclesiasticorum Latinorum, n° 32, Temsky, Vienne, 1896.
[38] H. Schmid, « Probabilistic Part-of-Speech Tagging Using Decision Trees », in *International Conference on New Methods in Language Processing* (Manchester), UMIST, 1994, p. 44-49.
[39] J. Schmitt (éd.), *Palladii Rutilii Tauri Aemiliani uiri inlustris Opus agriculturae*, Teubner, Leipzig, 1898.
[40] C. Silvi, « Citer Pline dans les encyclopédies médiévales : l'exemple des notices zoologiques chez Thomas de Cantimpré et Vincent de Beauvais », *Archives Internationales d'Histoire des Sciences* **61** (2011), n° 166-167, p. 27-55.
[41] M. Simard, G. F. Foster & P. Isabelle, « Using cognates to align sentences in bilingual corpora », in *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research : distributed computing-Volume 2*, IBM Press, 1993, p. 1071-1082.
[42] « SourcEncyMe (Sources des Encyclopédies Médiévales) », 2007, `http://sourcencyme.irht.cnrs.fr`.
[43] B. Van den Abeele, « Bestiaires encyclopédiques moralisés. Quelques succédanés de Thomas de Cantimpré et de Barthélemy l'Anglais », *Reinardus Yearbook of the International Reynard Society* **7** (1994), n° 1, p. 209-228.
[44] ———, « Vincent de Beauvais naturaliste : les sources des livres d'animaux du *Speculum naturale* », in *Lector et compilator : Vincent de Beauvais, frère prêcheur : un intellectuel et son milieu au* XIII^E SIÈCLE (S. Lusignan, M. Paulmier-Foucart & M.-C. Duchenne, éds.), Créaphis, Grâne, 1997, p. 127-151.
[45] A. Van Oppenraaij (éd.), *Aristotle De Animalibus, Michael Scot's Arabic-Latin Translation, Part Three : Books XV-XIX : Generation of Animals*, Brill, Leiden, Boston et Cologne, 1992.
[46] ——— (éd.), *Aristotle De Animalibus, Michael Scot's Arabic-Latin Translation, Part Two : Books XI-XIV : Parts of Animals*, Brill, Leiden, Boston et Cologne, 1998.
[47] « Transcription de la traduction de l'Historia animalium d'Aristote par Michel Scot ».
[48] M. J. Wise, « YAP3 : Improved detection of similarities in computer program and other texts », *ACM SIGCSE Bulletin* **28** (1996), n° 1, p. 130-134.
[49] K. Zangemeister (éd.), *Pauli Orosii historiarum adversum paganos libri VII*, Bibliotheca scriptorum Graecorum et Romanorum Teubneriana, Teubner, Leipzig, 1889.

Abstract. — With his encyclopaedia *Speculum maius*, the XIIIth century Dominican Vincent de Beauvais tries to form a general knowledge synthesis. To do so, he gathers information coming from a multitude of different sources, christian as well as pagan, from classical Antiquity as well as from Middles Ages. Most of his work's notices contain an explicit mention of the sources from which they were drawn, unlike many other medieval encyclopaedias. This feature allows using the *Speculum maius* as an experimentation dataset, and applying supervised learning and text mining techniques in order to automatically link the encyclopaedic notices to their sources. In this paper, we undertake such an exercise for the zoological books of the encyclopaedia, and we analyze the contributions, limitations and perspectives of the results we have obtained, having in mind to apply our methods to encyclopaedias which do not mention their sources in the future.

Keywords. — Text Mining, Text Reuse, Medieval History, Vincent de Beauvais, Medieval Encyclopaedia.

Resumen. — Con su enciclopedia llamada *Speculum maius*, el dominicano del siglo XIII Vincent de Beauvais intenta constituir una síntesis general del conocimiento. Para ello, reúne información de una multitud de fuentes diferentes, cristianas y paganas, antiguas y medievales. La mayoría de las fichas de su obra contienen una mención explícita de las fuentes en las que se inspiran, a diferencia de muchas enciclopedias medievales. Esta característica permite utilizar el *Speculum maius* como base para la experimentación, y aplicarle técnicas de aprendizaje supervisado y de minería de textos con el objetivo de vincular automáticamente los registros enciclopédicos con sus fuentes. En el presente artículo nos dedicamos a este ejercicio para los libros de zoología de esta enciclopedia y analizamos posteriormente los aportes, límites y perspectivas de los resultados obtenidos con miras a su futura aplicación a otras enciclopedias cuyos registros no mencionan sus fuentes.

Palabras claves. — Minería de textos, reutilización de texto, historia medieval, Vincent de Beauvais, enciclopedia medieval.

Manuscrit reçu le 19 août 2018, accepté le 16 mars 2019.