“ Bifidité ” et évolution philologie computationnelle des textes en langue d’oïl

Jean-Baptiste Camps

« Bifidité » et évolution

philologie computationnelle des textes en langue d'oïl

Un siècle, parsemé de vifs débats, nous sépare de l'observation de Joseph Bédier sur une propriété saillante des formes des stemmata, leur « bifidité ». Sur cette observation, un débat fondamental s'est construit, opposant la thèse bédiériste d'un biais méthodologique à celle d'un résultat du processus de transmission des textes. Cet article, en se concentrant sur le domaine de langue d'oïl, reviendra sur cette question d'« évolution » des traditions textuelles, dont on montrera qu'elle rejoint un problème plus général propre à l'ensemble des arbres tracés par les disciplines évolutionnistes, celui dit des « arbres déséquilibrés », bien connu dans le paradigme darwinien. On montrera ainsi comment la philologie peut se comprendre comme une science de l'évolution des textes, et que ses problèmes fondamentaux, tels que celui qui vient d'être énoncé, peuvent être avec profit abordé avec l'appui des méthodes computationnelles. Ces méthodes peuvent être mises au service du test d'hypothèse, en conjoignant démarche de modélisation et analyse de données, et en tirant profit de l'interdisciplinarité. Les ressources de l'intelligence artificielle peuvent en outre être mobilisées pour fonder ces approches sur des corpus de plus en plus vastes, favorisant des études à des échelles peu envisagées jusqu'ici.

I. La philologie (computationnelle) comme science des textes et de leur évolution

Il est établi que la philologie en tant que science a émergé dans un mouvement concomittant avec d'autres disciplines, au premier chef la biologie et la linguistique diachronique, avec lesquelles elle partage un paradigme fondé sur l'héritage et l'innovation, et une métaphore arborée comme mode de représentation (stemma, phylogénie et `*Stammbaum*' des langues)[1]. Dans la lignée de cette origine remontant aux XVIIIe et XIXe siècles, j'arguerai que la philologie, même si elle a pu également être définie depuis comme « l'art d'éditer » ou « l'art de lire » les textes[2], peut encore être posée comme une science de l'évolution, dont l'objet principal est constitué par les textes (et par eux, la culture), ce qui la rapproche tout en la distinguant d'autres disciplines évolutionnistes, telles que biologie, linguistique ou anthropologie[3].

Ainsi, depuis Bengel (pour le principe) et Schlyter (pour sa réalisation), la philologie repose sur un modèle arboré comme mode de représentation de la transmission et de l'évolution des textes, à tel point que, comme le relève Froger, on ferait mieux de parler de « méthode schlytérienne » plutôt

1 Sur ce dernier point, voir notamment C. GINZBURG, *Somiglianze di famiglia e alberi genealogici. Due metafore cognitive*, in C. C. HÄRLE, *Ai limiti dell'immagine (Estetica e critica)*, Macerata, Quodlibet, p. 1000-1024.

2 F. DUVAL, *Les mots de l'édition de textes*, Paris, École nationale des chartes, 2015 ; J. MONFRIN, *Leçon d'ouverture du cours de philologie romane à l'Ecole des chartes*, in «Bibl. de l'École des Chartes» 116 (1958), p. 170–193.

3 Pour une comparaison entre méthodes philologique et autres disciplines de l'évolution, on se reportera à A. HOENEN, *8 Evolutionary models in other disciplines*, in *Handbook of Stemmatology: History, Methodology, Digital Approaches*, Berlin, De Gruyter, 2020, p. 534–586. On pourra aussi se référer à l'emploi du terme, quoique sans doute avec un sens moins marqué, par Paul Zumthor, pour qui « dans son acception la plus générale, la philologie […] vise à saisir, dans leurs manifestations linguistiques, le génie propre d'un peuple ou d'une civilisation et leur évolution culturelle » ; P. ZUMTHOR, *Philologie*, in *Encyclopædia Universalis*, https://www.universalis.fr/encyclopedie/philologie/.

que lachmanienne[4]. Schlyter a en effet tracé un arbre qui, non content d'être le tout premier stemma connu, se distingue par son niveau d'aboutissement, plus moderne non seulement que ceux de ses successeurs immédiats, mais aussi que la plupart de ceux qui seront produits au XIX[e] siècle. Mais il y a plus, car le stemma de Schlyter remplit également les critères lui permettant de concourir au titre de tout premier arbre évolutif darwinien : représentation des nœuds ancestraux et de leurs descendants, nœuds ancestraux représentant des objets de même nature que les nœuds descendants (i.e., ils représentent des maillons supposés de la transmission, pas des groupes ou des classes de témoins, comme le feraient les regroupements de qualité supposée ou d'origine géographique que l'on trouve parfois chez certains pionniers), et séparation de différentes lignées[5].

En réalité, au-delà de la représentation arborée, Schlyter inaugure également, et sans le dire explicitement, le concept-clé qu'est l'utilisation des innovations ou « fautes communes » comme base de l'établissement de la généalogie[6]. Conceptuellement, cette approche dite des « fautes communes », précisée ensuite par des philologues tels que Paul Lejay[7], est identique au principe de « *descent with modification* »[8] (« descendance avec modification »), qui est la définition darwinienne du processus d'évolution.

L'emploi de la métaphore de l'arbre et du concept de descendance avec modification, c'est-à-dire d'évolution, suffisent à caractériser la philologie comme une science évolutionniste. Mais en outre, durant son histoire, la philologie est progressivement passée d'une quête du seul texte original pour s'enrichir d'un intérêt pour l'histoire des textes, de leurs transformations et acclimatations à des périodes et niches géographiques diverses, autrement dit d'un intérêt pour leur processus d'évolution en lui-même (et ce qu'il révèle de la culture). On est ainsi passés de la formule de Gaston Paris, pour qui « la critique des textes a pour but de retrouver, autant que possible, la forme que l'ouvrage auquel elle s'applique avait en sortant des mains de l'auteur » à celle de Pasquali, qui se propose de « seguire la storia di un testo per tutto il mondo medievale, (…) vederlo diffondersi da una provincia all'altra, e anche calcolare a un dipresso come e quanto si sia colorato nelle sue peregrinazioni »[9].

---

4 J. A. BENGEL, *Apparatus criticus ad Novum Testamentum*, Tubingae, 1763; D. C. J. SCHLYTER – D. H. S. COLLINS, *Corpus iuris Sueo-Gothorum antiqui…*, Stockholm, 1827; J. FROGER, [Compte-rendu de:] *Westgöta-Lagen*, ed. par H. S. COLLIN et C. J. SCHLYTER, *Facsimile edition with an addendum by* Otto VON FRIESEN, in «Scriptorium» 32 (1978), p. 183–188.

5 D. MORRISON, *The first Darwinian evolutionary tree*, in *The Genealogical World of Phylogenetic Networks* (blog) 24 giugno 2013, http://phylonetworks.blogspot.com/2013/06/the-first-darwinian-evolutionary-tree.html.

6 FROGER, op. cit.

7 BENGEL, op. cit., écrivait déjà en 1763 que « *Codices duo pluresve, qui in libris & capitibus quibusdam singulariter congruunt (…) fere ab una stirpe communi descendunt, et fere pro uno codice numerari possunt* », mais c'est Lejay en 1888 qui l'énonce le plus clairement en disant qu'une « famille de manuscrits est constitué par leurs fautes communes, ou, si l'on préfère ce terme plus exact, par leurs innovations communes »; P. LEJAY, [Compte-rendu de] *Aeli Donati quod fertur Commentum Terenti... Recensuit* Paulus WESSNER, in «Rev. Crit. Hist. Litt.» 56 (1903), p. 168–172.

8 C. DARWIN, *On the Origin of species by means of natural selection*, London, John Murray, 1859, p. 420 : « the natural system is founded on descent with modification ; […] the characters which naturalists consider as showing true affinity between any two or more species, are those which have been inherited from a common parent, and, in so far, all true classification is genealogical ».

9 G. PARIS – L. PANNIER (Edd.), *La vie de saint Alexis: poème du XI[e] siècle et renouvellements des XII[e], XIII[e], et XIV[e] siècles publ. avec préfaces, variantes, notes et glossaires,* Paris, A. Franck, 1872; G. PASQUALI, *Paleografia quale scienza dello spirito in Nuova Antologia, 1 giugno 1931*, in *Pagine stravaganti di un filologo*, Lanciano, Giuseppe Carabba, 1934, p. 181–205.

Au-delà d'une tradition rhétorique voulant commencer tout texte par une définition des termes du sujet, le fait de reposer la philologie comme science de l'évolution permet de profiter d'un bagage conceptuel et méthologique interdisciplinaire, qui a des applications intéressantes à la philologie. Par exemple, le problème de la « bifidité » des stemmata, posé par Bédier il y a presque 100 ans, et qui sera traité dans les prochaines sections, peut-être rattaché au problème plus général des « arbres d'évolution déséquilibrés » (ou asymétriques). Plus largement, tout une gamme de questions de (macro)évolution culturelle s'ouvrent ainsi, telles que :

— Pourquoi certaines œuvres écrites ont-elles survécu, alors que d'autres ont disparu ?

— Quelle proportion avons-nous perdu de la culture et des documents du passé ?

— Quelle est la part du hasard (dérive) et de la sélection dans ce que nous avons conservé ou perdu ? dans la popularité des œuvres ? la constitution de vulgate et de canons littéraires ?

— Comment le contenu textuel ou le contexte jouent-ils un rôle dans la transformation ou la sélection des variantes, des versions ou des textes ?

— Comment la variété des textes a-t-elle augmenté ou diminué au fil du temps ? Existe-t-il des facteurs identifiables ?

— Comment peut-on décrire le processus menant à la fixation de variantes, d'une version vulgate ?

— Quel est l'effet de l'existence de différentes régions ou niches pour les textes, en particulier quel est précisément le rôle des régions plus isolées ou insulaires, que les philologues ont déjà souvent mis en évidence, surtout à la suite de Pasquali[10] ?

— Dans quels contextes se produisent des transmissions latérales (contaminations) ? Ou des évolutions convergentes, comme les variantes polygéniques ?

Ce n'est bien sûr là qu'un aperçu des questionnaires possibles. L'objet en est surtout de montrer qu'il est possible de réenvisager différemment des questionnements depuis longtemps présents dans la philologie, d'une manière notamment congruente avec l'emploi des méthodes computationnelles comme mode de test d'hypothèse, permettant de faire sens du déluge de données auquel nous sommes confrontés.

Les dernières décennies ont en effet été marqué par un « déluge de données »[11], qui ne cesse encore de s'amplifier. Celui-ci touche directement la disponibilité de nos sources, en particulier des manuscrits numérisés et des imprimés historiques. Sur la seule bibliothèque numérique *Gallica* de la Bibliothèque nationale de France, le nombre de documents disponibles est ainsi passé de quelques milliers à plus de 8 millions en vingt ans[12]. La croissance exponentielle du nombre de documents ne s'accompagne pas encore nécessairement d'une disponibilité systématique (ou d'une

10 G. PASQUALI, *Storia della tradizione e critica del testo*, Florence, F. le Monnier, 1934, a souligné l'importance de ce qu'il appelle les « zones périphériques » (*zone periferiche*) où les premiers stades d'une tradition sont mieux conservés, étant les dernières à être touchées par l'émergence de nouvelles versions de la vulgate rayonnant à partir des centres culturels (voir aussi P. TROVATO, *2 The genealogical method: 2.4 Neo-Lachmannism: A new synthesis?*, in *Handbook of Stemmatology*, cit., p. 109–138). On connaît notamment le rôle de l'Italie du Nord, et en particulier de la Vénétie, en tant que « area laterale e conservativa di frontiera » dans la préservation des textes et manuscrits des troubadours, selon G. FOLENA, *Tradizione e cultura trobadorica nelle corti e nelle città venete*, in *Culture et lingue nel Veneto medievale*, Padoue, Editoriale Programma, 1990, p. 1–37.

11 G. BELL – T. HEY – A. SZALAY, *Beyond the data deluge*, in «Science» 323 (2009), p. 1297–1298.

qualité satisfaisante) de la transcription automatique, mais nous sommes en passe d'y parvenir, grâce à des systèmes de reconnaissance de textes manuscrits de plus en plus généralistes et performants, et il est d'ores-et-déjà possible de constituer des corpus de centaines ou de milliers de manuscrits[13].

Qu'est-ce que cela change ? Tout d'abord, la large disponibilité et l'accès facile aux manuscrits numérisés - et à des transcriptions automatisées relativement fiables - ont, à mon avis, pour effet de revaloriser et de remettre en avant la dimension critique du travail philologique. En effet, quel est l'intérêt de continuer à produire des éditions comme transcription d'un unique témoin avec le moins de modifications possibles (comme le voudrait une certaine conception extrême du bédierisme ou de la « Nouvelle Philologie »), quand on peut accéder ou récolter des images numérisées de milliers de manuscrits, et en obtenir une version en texte intégral grâce à un processus reproductible et toujours amélioré de reconnaissance de texte manuscrit[14] ? Il en résulte un recentrage sur la dimension critique du travail philologique, qu'il s'agisse d'enrichir le document source de métadonnées, d'annotations et d'autres informations utiles à l'analyse, de fournir des éditions comme « hypothèses de travail »[15] et instrument de connaissance des traditions textuelles, ou d'aller au-delà des textes individuels pour suivre les évolutions à long terme ou grande échelle.

D'autre part, ce déluge de données appelle de nouvelles procédures analytiques, car il constitue également un risque, en fournissant finalement trop de données pour être réellement interprétées par les méthodes traditionnelles. L'analyse de données et la « lecture distante » (pour reprendre le concept de Franco Moretti[16]) sont bien sûr des éléments de solution pour gérer des quantités de plus en plus importantes d'informations. Mais dans son expression la plus naïve, qui consisterait à laisser le sens émerger quasi spontanément des données, il s'agit d'une illusion. Au contraire, une plus grande quantité d'informations nécessite des schémas explicatifs plus compacts et plus généraux. En d'autres termes, nous devrions nous efforcer de trouver ce qui a une signification générale, et pas seulement ce qui pourrait être irréductiblement singulier dans chaque document.

Pour cette raison, le plan de cet article, qui essaie de conjoindre réflexion méthodologique, théorie, modélisation et analyse de données, tout en servant de démonstration sur le cas particulier de la « bifidité », part du plus abstrait (la théorie et les hypothèses à tester) pour aller vers le concret (les données empiriques).

12 Sur ce sujet, on pourra aussi se reporter à mon article, J.-B. CAMPS, *La philologie computationnelle à l'École des chartes: premier bilan et perspectives*, in «Bibl. de l'École des Chartes» 176 (2021), p. 193-216.

13 À titre d'exemple, le *Corpus of Medieval French Epics and Romances* comporte actuellement la transcription automatique de 409 témoins, pour presque 40 millions de mots ; voir J.-B. CAMPS et al., *Make Love or War? Monitoring the Thematic Evolution of Medieval French Narratives*, in A. ŠEĻA – F. JANNIDIS – I. ROMANOWSKA (Edd.), *Proceedings of the Computational Humanities Research Conference 2023 Paris, France, December 6-8, 2023*, Paris, 2023 (CEUR workshop proceedings, 3558), p. 734-756 <https://ceur-ws.org/Vol-3558/>.

14 Voir les points de vue de A. CORBELLARI, *Le texte médiéval. La littérature du Moyen Age entre topos et création*, in «Poétique» 163 (2010), p. 259–273, et le tout à fait visionnaire E. HICKS, *Éloge de la machine: transcription, édition, génération de textes*, in «Romania» 103 (1982), p. 88–107.

15 G. CONTINI, *Ricordo di Joseph Bédier*, in «Letteratura» 3 (1939), p. 145.

16 F. MORETTI, *Distant reading*, Londres, Verso, 2013.

II. Modèles pour la transmission des textes

Le problème des arbres déséquilibrés

Il n'est plus la peine de présenter aux philologues romans les observations faites par Bédier, il y a plus d'un siècle, sur la particularité des formes des stemmata, et leur forte propension à ce qu'il appelle la « bifidité »[17]. La tendance des racines de ces arbres à présenter un degré de sortie égal à deux a été, depuis Bédier, confirmée par d'autres enquêtes, quoi qu'à des niveaux variables (entre environ 60 et 90%) et généralement inférieurs à l'estimation qu'il avait faite (95,5% ; table 1) ; notons néanmoins que ces enquêtes se limitent aux traditions pour lesquelles au moins un stemma a été proposé (ce qui exclut les textes en témoin unique, et les traditions qui se prêtent mal à la réalisation d'un stemma).

Ce qui a été moins souvent noté, en revanche, c'est que cette « bifidité » n'est pas la seule propriété saillante de la forme de ces graphes. L'extrême rareté des liens directs entre témoins (déjà notée par Gaston Paris[18]), le caractère localisé de la tradition survivante (remontant souvent à un archétype plutôt qu'à l'original), et surtout, la tendance généralisée à l'asymétrie des branches, sont des propriétés tout aussi intéressantes et potentiellement révélatrices des mêmes paramètres propres à la transmission des textes.

% bifid.	N	Lang.	Genre	Statut	Collection	Réf.
95,50	110	Fr, Lat, Eng, Ger		*Non précisé*		BÉDIER, op. cit
69,00	130	Pro	lyrique troub.	*Non précisé*		SHEPARD, op. cit.
75,50	94	Fr		Tous	Rép. Bossuat	CASTELLANI, op. cit.
82,5	86	Fr		Définitifs seuls	Rép. Bossuat	CASTELLANI, op. cit.
83,1	89	Norrois		Définitifs seuls	*Bibl.* et *Ed. Arnamagnaena*	HAUGEN, op. cit.
90	30	Fr	gestes	Tous	*Open Stemmata*	CAMPS, GABAY, RIVA, op. cit.
68,2	22	Fr	romans	Tous	*Open Stemmata*	CAMPS, GABAY, RIVA, op. cit.

Table 1 : pourcentage de bifidité dans des collections de *N* stemmata, dans différentes langues, parfois sélectionnés selon des critères de genre littéraire, de statut définitif ou provisoire des stemmata, ou selon une collection donnée.

Ce qui a été encore moins noté (ou plutôt, à ma connaissance, jamais noté) est que ce problème des « arbres déséquilibrés » (ou asymétriques) n'est pas du tout unique à la philologie et se présente généralement dans toutes les disciplines qui utilisent des arbres d'évolution, telles que biologie et linguistique (Figure 1). En biologie, il a été noté par un contemporain de Bédier, G. Udny Yule, qui constatait déjà en 1925, que la distribution du nombre d'espèces par *genus* était « à longue traîne », avec un très grand nombre de *genus* ne possédant qu'une seule espèce survivante, et un tout petit nombre de *genus* ayant un très grand nombre d'espèces[19]. Chez les serpents, par exemple, il relève que 131 *genus* ne présentent qu'une seule espèce, tandis que 4 en possèdent plus de 35. Il montre

17 Sur ce sujet, voir en dernier lieu F. DUVAL, *La `Tradition manuscrite du Lai de l'Ombre' de Joseph Bédier ou la critique textuelle en question*, Paris, Honoré Champion, 2021.

18 PARIS – PANNIER, op. cit., p. 10.

19 G. U. YULE, *A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S*, in «Philos. Trans. R. Soc. Lond. Ser. B» 213 (1925), p. 21–87.

également qu'un processus aléatoire de naissance et de tirage au sort permet d'expliquer assez bien cette distribution. Aldous a plus récemment explicité que le modèle proposé par Yule se traduit par des arbres phylogénétiques aux formes très asymétriques, ou « déséquilibrées » (*imbalanced*)[20]. L'intuition en est assez simple à comprendre : une branche ayant déjà de nombreuses ramifications à plus de chances d'en voir de nouvelles surgir qu'un rameau isolé, toutes choses égales par ailleurs.

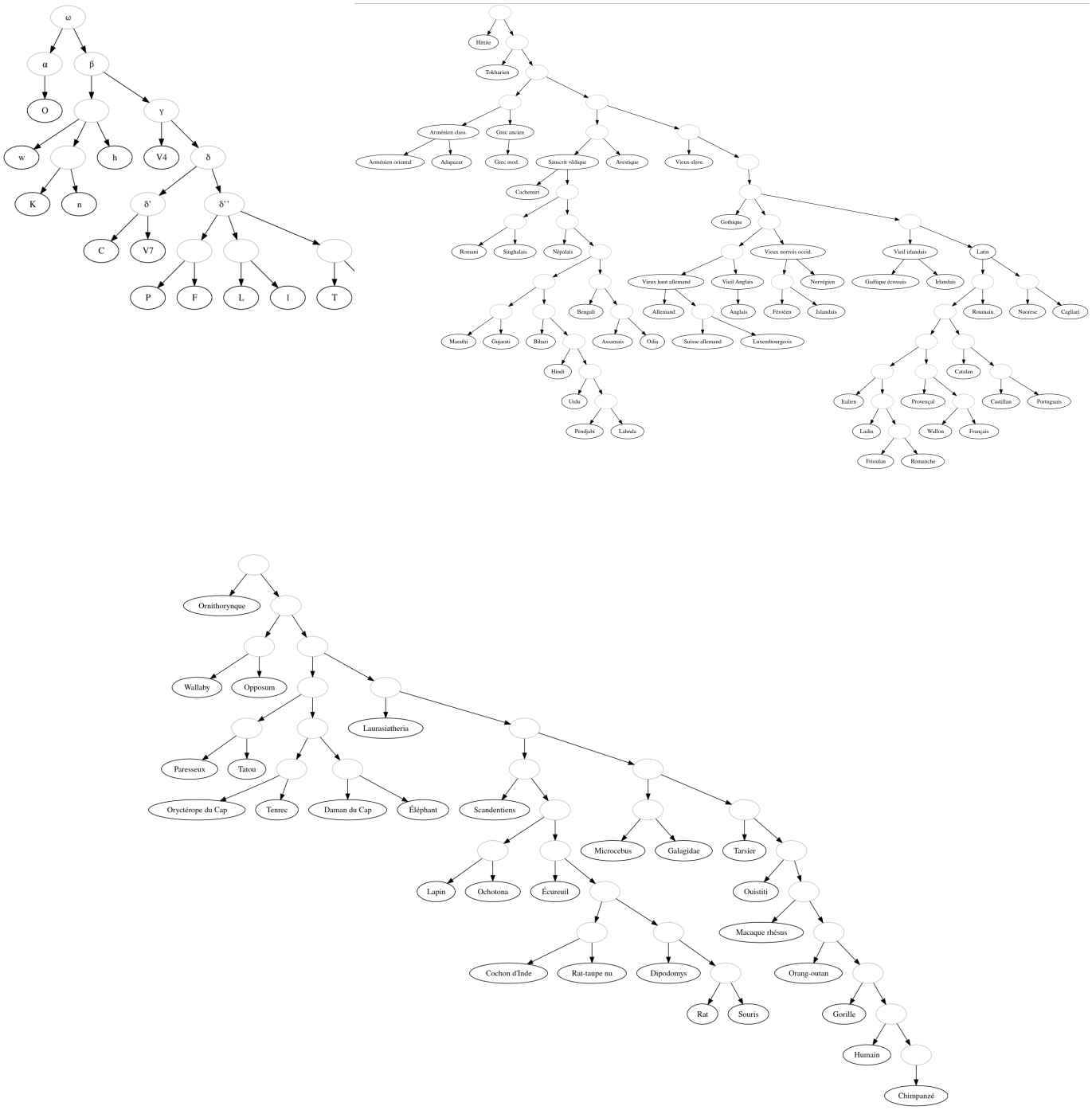

Figure 1 : arbres d'évolution asymétriques en philologie (*Roland*), biologie (mamifères) et linguistique (langues indo-européennes). D'après : C. SEGRE, *La chanson de Roland*, Milan et Naples, R. Ricciardi, 1971 (= Documenti di filologia, 16) ; J. E. TARVER, et al., *The interrelationships of placental mammals and the limits of phylogenetic inference*, «Genome Biology

20 D. J. ALDOUS, *Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today*, in «Stat. Sci.» 16 (2001), p. 23–34.

and Evolution», 8 (2016), p. 330–344 ; W. CHANG, et al., *Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis*, «Language», 91 (2015), pp. 194–244.

Chez les biologistes, comme chez les philologues, ce problème des arbres d'évolution déséquilibrés a causé un large débat, dont les deux thèses opposées seraient, d'une part, que ces formes sont dues à un biais méthodologique[21], et, d'autre part, qu'elles sont au contraire révélatrices du processus d'évolution à l'œuvre. Si l'on accepte cette seconde thèse, la question se pose alors de savoir si ces formes observées correspondent à un processus complètement aléatoire, ou si des déterminismes en sont la cause (via la sélection naturelle, les radiations évolutives, les extinctions de masse…) ?

Le *status quaestionis* actuel est en effet qu'un processus purement stochastique de naissance et de mort mène à des arbres très déséquilibrés, mais que les arbres phylogénétiques reconstruits à partir de données réelles sont encore plus déséquilibrés qu'ils ne le devraient sous l'effet du seul hasard[22]. Depuis une réunion fondatrice de paléontologues et biologistes des populations à Woods Hole au début des années 1970, tout un champ de recherche s'est ainsi ouvert liant la question de la topologie des arbres (et notamment de leur caractère déséquilibré), avec l'étude des processus de macroévolution, des taux d'apparition et d'extinction d'espèces, et du rôle de la biogéographie[23]. La question des modèles mathématiques d'évolution et de l'analyse statistique des propriétés des arbres y jouent un rôle important, en dépit des réticences que pourraient avoir eues dans un premier temps les biologistes « non-mathématiciens », encore attachés à l'idée que la « *macroevolutionary history, like human history, is a mosaic of singular events not amenable to mathematical modeling* »[24].

Chez les philologues, parmi les trois hypothèses possibles (artefact méthodologique, processus déterministe ou aléatoire), et d'ailleurs non mutuellement exclusives, la première a concentré une grande partie de l'attention, sans être réellement démontrée. Bédier lui-même, dont on sait qu'il fut un tenant acharné de la première hypothèse, feint un instant d'envisager ces alternatives, pour les écarter d'emblée par une pirouette rhétorique :

> Un fait de bipartition dans l'histoire de la transmission des textes n'étonne pas, mais la loi de bipartition étonne, et l'étonnement croît à mesure qu'elle se fait plus impérieuse et qu'elle affecte davantage la constance, la majesté, la nécessité d'une loi de la nature. Nécessaire et par là même absurde, il ne se peut pas qu'elle ait régi les destinées des textes au cours des siècles anciens[25]

« Nécessaire et par là même absurde, il ne se peut pas… », et pourquoi donc ? D'autres philologues ont heureusement pris le relais de Bédier, pour explorer différentes perspectives, d'une manière parallèle quoique généralement indépendante de ce qui a pu exister en biologie. Des approches descriptives, fondées sur l'analyse de collection de stemmata, ont amené comme on l'a vu

21 Voir par exemple E. STAM, *Does Imbalance in Phylogenies reflect only bias?*, in «Evolution» 56 (2002), p. 1292–1295, qui n'observe pas de corrélation entre la qualité des données utilisées et le degré d'équilibre des arbres. Voir aussi A. O. MOOERS – S. B. HEARD, *Inferring Evolutionary Process from Phylogenetic Tree Shape*, in «Q. Rev. Biol.» (1997), p. 31-54, part. p. 38-41, « Nonevolutionary Explanations for Imbalance ».

22 ALDOUS, op. cit.

23 MOOERS – HEARD, op. cit.

24 ALDOUS, op. cit., p. 31.

25 J. BÉDIER, *La tradition manuscrite du Lai de l'ombre : réflexions sur l'art d'éditer les anciens textes*, in «Romania» 54 (1928), p. 161–196 et 321–356, à la p. 172.

à des estimations légèrement inférieures, entre 68 et 90% de bifidité (95,5% chez Bédier)[26]. Pour l'avenir, la constitution de bases de données ouvertes et pérennes de stemmata fournissent un moyen d'élargir ces enquêtes, en les rendant plus reproductibles que la papier-calque des pionniers[27].

Des approches d'ordre combinatoire ont cherché à savoir, pour un nombre donnée de témoins, quels étaient l'ensemble des topologies possibles (et leur taux de bifidité). Depuis Maas, ces approches ont étudié le nombre de configurations possibles pour un nombre donné de nœuds, en considérant les tiges comme des arbres aléatoires, des graphes acycliques dirigés ou en les formalisant comme des « arbres de Cayley » ou des « arbres de Greg » (nommés d'après Walter Wilson Greg) et en tenant compte ou non de la transmission latérale, c'est-à-dire de la contamination[28]. Bien que nécessaires, ces recherches ont le défaut de considérer toutes les topologies comme équiprobables (ce que rien ne garantit). Elles pourraient toutefois être mises à profit pour mesurer des biais méthodologiques, car elles donnent une idée de la topologie qu'auraient les stemmata, toutes choses égales par ailleurs, dans le cas extrême où toutes les hypothèses philologiques seraient d'égale valeur et où le choix des philologues équivaudrait à un tirage aléatoire[29].

Le troisième type d'approche s'est penché sur le rôle des pertes de manuscrits, de la décimation (les biologistes parleraient d'extinction) sur la topologie des arbres. L'idée en remointe au moins à Walter W. Greg, voire même à Gaston Paris[30]. De cette idée découle une série d'approches probabilistes, basées sur l'application *a posteriori* de la décimation à un arbre statique et déjà entièrement constitué, généralement sous la forme d'une probabilité de destruction uniforme[31].

---

26 W. P. SHEPARD, *Recent theories of textual criticism*, in «Mod. Philol.» 28 (1930), p. 129–141; A. CASTELLANI, *Bédier avait-il raison? La méthode de Lachmann dans les éditions de textes du Moyen Age*, Fribourg, Éditions Universitaires, 1957; O. E. HAUGEN, *The silva portentosa of stemmatology: Bifurcation in the recension of Old Norse manuscripts*, in «Digit. Scholarsh. Humanit.» 31 (2015), p. 594–610. Pour des travaux similaires en biologie, voir par exemple M. G. B. BLUM – O. FRANÇOIS, *Which Random Processes Describe the Tree of Life? A Large-Scale Study of Phylogenetic Tree Imbalance*, in «Syst. Biol.» 55 (2006), p. 685–691.

27 J.-B. CAMPS – S. GABAY – G. F. RIVA, *Open Stemmata: A Digital Collection of Textual Genealogies*, in *EADH2021: Interdisciplinary Perspectives on Data, 2nd International Conference of the European Association for Digital Humanities*, Krasnoyarsk, 2021. Pour un précédent concernant les arbres phylogénétiques, voir W. H. PIEL et al., *TreeBASE: a database of phylogenetic information*, in *Proceedings of the 2nd International Workshop of Species*, Tsukuba, National Institute for Environmental Studies, 2000, p. 41–47.

28 P. MAAS, *Leitfehler und stemmatische Typen*, in «Byzantinische Z.» 37 (1937), p. 289–294 ; W. W. GREG, *The calculus of variants: an essay on textual criticism*, Oxford, Clarendon Press, 1927 ; C. FLIGHT, *How Many Stemmata ?*, in «Manuscripta» 34 (1990), p. 122–128 ; W. HERING, *Zweispaltige stemmata: Zur Theorie der textkritischen Methode*, in «Philologus» 111 (1967), p. 170–185 ; A. HOENEN – S. EGER – R. GEHRKE, *How Many Stemmata with Root Degree k?*, in *Proceedings of the 15th Meeting on the Mathematics of Language*, London, Association for Computational Linguistics, 2017, p. 11-21 ; A. HOENEN, *An open problem in computational stemmatology - a model for contamination*, in «Um. Digit.» 5 (2019), doi:10.6092/issn.2532-8816/8555 ; M. JOSUAT-VERGÈS, *Derivatives of the tree function*, in «The Ramanujan Journal» 38 (2015), p. 1–15; D. NAJOCK – C. C. HEYDE, *On the number of terminal vertices in certain random trees with an application to stemma construction in philology*, in «J. Appl. Probab.» (1982), p. 675–680.

29 Pour des travaux de même nature en biologie, voir les modèles dits « PDA » et « EPT » (A. O. MOOERS – S. B. HEARD, op. cit., p. 36-37).

30 GREG, *Recent theories of textual criticism*, in «Mod. Philol.» 28 (1931), p. 401–404 ; PARIS – PANNIER, op. cit., p. 10, écrivait déjà « L'immensité des pertes que nous avons faites en fait de manuscrits du moyen-âge n'apparaît nulle part avec plus d'évidence que lorsqu'on possède plusieurs textes d'un même ouvrage. En effet, il est infiniment rare que l'un de ces textes soit copié sur l'autre; presque toujours ce sont les extrêmes pousses de branches parfaitement distinctes, qui forment autour de la tige, qu'on ne possède jamais, une vaste ramification ».

31 J. FOURQUET, *Le paradoxe de Bédier*, in «Mélanges 1945, II : Études littéraires», fasc. 105 , Strasbourg, Publ. de la Faculté des lettres de l'Université de Strasbourg ; F. WHITEHEAD – C. E. PICKFORD, *The two-branch stemma*, in «Bull. Bibliogr. Société Int. Arthurienne» 3 (1951), p. 83–90; CASTELLANI, op. cit.; V. A. DEARING, *Methods of Textual Editing*, University of California, William Andrews Clark Memorial Library, 1962; W. HERING, op. cit. ; A.

Kleinlogel y ajoute la prise en compte de l'âge pour attribuer une probabilité de survie plus élevée aux manuscrits les plus jeunes, tandis que Guidi et Trovato ont introduit l'utilisation d'arbres réels issus de la tradition des éditions imprimées[32]. Si elles apportent un éclairage appréciable sur le lien entre la décimation et la forme des stemmata, ces études supposent que la décimation est non biaisée et distribuée de manière homogène, et ne tiennent toujours pas compte de la nature dynamique de la transmission textuelle, où les copies sont simultanément faites et perdues.

La dernière catégorie d'approche, encore très peu explorée par les philologues[33], tient à l'utilisation de modèles stochastiques permettant de modéliser et simuler des traditions manuscrites, pour explorer le rôle du hasard dans les formes que nous observons, et les écarts éventuels par rapport à celui-ci. L'utilisation de processus dits « de naissance et de mort » et de simulations informatiques, après des explorations initiales par Haigh (qui ne l'a examiné que comme un processus de naissance pur) et Dearing, a été poursuivie par Weitzman dans deux articles fondamentaux, bien qu'insuffisamment appréciés[34]. En modélisant la transmission textuelle comme un processus de naissance et de mort, en effectuant des simulations et en prenant des études de cas dans la littérature classique grecque et latine, Weitzman a inauguré une méthodologie potentiellement féconde, bien qu'elle n'ait suscité en fait presque aucun successeur et que la recherche dans ce domaine soit restée pour l'essentiel dans l'état où il l'avait laissée en 1987[35]. Sans que le lien direct n'ait été bien explicité jusqu'à présent, les travaux de Cisne partagent des similarités théoriques avec ceux de Weitzman : Cisne a en effet utilisé les méthodes de la paléodémographie et utilisé des modèles de dynamique de population, de paramètres similaires à ceux de Weitzman, pour les ajuster aux distributions d'âge observées pour quelques œuvres médiévales[36]. De ce travail très novateur, Cisne tire des estimations des pertes de manuscrits, probablement fortement sous-estimées en raison de l'utilisation de données discutables et de certaines hypothèses du modèle, principalement la nature constante des paramètres, qui ignoraient

KLEINLOGEL, *Das Stemmaproblem*, in «Philologus-Zeitschrift für antike Literatur und ihre Rezeption» 112 (1968), p. 63–82.

32 A. KLEINLOGEL, op. cit. ; V. GUIDI – P. TROVATO, *Sugli stemmi bipartiti. Decimazione, asimmetria e calcolo delle probabilita*, in «Filol. Ital.» I (2004), p. 9–48.

33 Elle l'a été considérablement plus en biologie, depuis les travaux fondateurs de YULE, op. cit. Les modèles stochastiques ont ainsi été au cœur des travaux du « groupe de Woods Hole », et continuent d'être explorées aujourd'hui ; MOOERS – HEARD, op. cit ; F. BIENVENU – F. DÉBARRE – A. LAMBERT, *The split-and-drift random graph, a null model for speciation*, in «Stochastic Processes and their Applications» 129 (2019), p. 2010–2048.

34 J. HAIGH, *The manuscript linkage problem*, in F. R. HODSON (Ed.), *Mathematics in the Archaeological and Historical Sciences*, Edinburgh, Edinburgh University Press, 1971, p. 396–400; V. A. DEARING, *Principles and Practice of Textual Analysis*, University of California Press, 1974; M. P. WEITZMAN, *Computer simulation of the development of manuscript traditions*., in «ALLC Bull.» 10 (1982), p. 55–59; WEITZMAN, *The Evolution of Manuscript Traditions*, in «J. R. Stat. Soc. Ser. Gen.» 150 (1987), p. 287–308.

35 À ma connaissance, seuls trois articles peuvent être considérés comme étroitement liés au travail de Weitzman, sans être nécessairement des suites directes. H. K. GJESSING – R. H. PIERCE, *A stochastic model for the presence/absence of readings in Niðrstigningar Saga*, in «World Archaeol.» 26 (1994), p. 268–294 proposent un modèle incluant une représentation des variations dans le texte lui-même (les variantes), tandis que P. CANETTIERI et al., *Philology and Information Theory*, in «Cognitive Philolology» 1 (2008), réfléchissant sur le sujet, l'ont comparé à la « ruine du joueur » appliquée par Raup (D. M. RAUP, *Extinction: Bad Genes or Bad Luck?*, New York, W.W. Norton & Company, 1992) à l'extinction des espèces. A. HOENEN, *Silva Portentosissima – Computer-Assisted Reflections on Bifurcativity in Stemmas*, in *Digital Humanities 2016: Conference Abstracts*, Cracovie, Jagiellonian University & Pedagogical University Kraków, 2016, a réalisé quelques expériences sur la simulation de traditions où la probabilité de perte était dépendante de l'âge et du degré de sortie de chaque manuscrit.

36 J. L. CISNE, *How Science Survived: Medieval Manuscripts' «Demography» and Classic Texts' Extinction*, in «Science» 307 (2005), p. 1305–1307.

la variation dans le temps[37]. Les recherches très récentes de Kestemont, Karsdorp *et al.*, qui ont poursuivi l'enquête sur la question des pertes, adoptent quant à elles une perspective différente, inspirée des études de biodiversité, pour estimer les pertes des romans de chevalerie médiévaux à travers l'Europe[38]. Quoique non sans lien avec ces travaux précédents, leur perspective demeure essentiellement synchronique, et ne se penche pas sur la question de la macroévolution et la forme des arbres ou les dynamiques de population.

Ainsi, les idées de Weitzman, bien qu'elles constituent la base probable sur laquelle construire de futures tentatives de modélisation, ont été dans l'ensemble insuffisamment étudiées et de manière peu concluante - si tant est qu'elles l'aient été. Pourtant, il existe un état de l'art en constante progression du côté de la biologie de l'évolution, qui ouvre de nombreuses perspectives, et articule considérations théoriques, modélisation mathématique et observations empiriques. En particulier, l'étude de la macroévolution s'est penchée sur des questions telles que les dynamiques de spéciation et d'extinction, et leur part d'aléatoire et de déterminisme, qui se manifestent dans la forme des phylogénies et la perte de branches de l'arbre de la vie[39] ; ou bien encore sur la variation dans le temps et l'espace, étudiés en macroécologie et en biogéographie[40]. Ces préoccupations rejoignent des questions qui préocuppent de longue date les philologues, telles que l'estimation des pertes, en termes de manuscrits et d'œuvres, ou bien encore le rôle des aires latérales ou insulaires dans la conservation. Enfin, l'étude du rôle du hasard nécessite l'établissement d'un modèle nul, c'est-à-dire rendant compte d'un processus tel qu'il serait s'il était entièrement aléatoire ; là aussi, un état de l'art significatif a été produit[41].

Une double méthodologie : modélisation stochastique et statistiques descriptives

Sur les bases qui viennent d'être énoncées, une méthodologie double se dessine. Il importe ainsi, d'une part, de concevoir un modèle nul, rendant compte du processus de transmission des textes par réplication manuscrite, et de l'autre, de collecter et analyser des données concernant des traditions réelles, leurs généalogies, et les propriétés principales des œuvres, versions et témoins. Une confrontation entre modèles et données observées peut ainsi servir de fondement à la caractérisation de l'évolution des textes, et à l'estimation, notamment, des pertes, du rôle du hasard ou des divers déterminismes envisageables, ou encore de la variation dans le temps et l'espace.

Dans un article récent, nous avons ainsi entrepris la conception d'un modèle nul de transmission des manuscrits[42]. Nous utilisons un processus stochastique de naissance et de mort, avec les paramètres λ (probabilité d'être copié pour un agent), μ (probabilité d'être détruit), T (nombre total

37 G. DECLERCQ, *Comment on «How Science Survived…»*, in «Science» 310 (2005), p. 1618–1618; J. L. CISNE, *Response to Comment on «How Science Survived…»*, in «Science» 310 (2005), p. 1618–1618.

38 M. KESTEMONT et al., *Forgotten books: The application of unseen species models to the survival of culture*, in «Science» 375 (2022), p. 765–769.

39 K. YESSOUFOU – T. J. DAVIES, *Reconsidering the Loss of Evolutionary History: How Does Non-random Extinction Prune the Tree-of-Life?*, in R. PELLENS – P. GRANDCOLAS (Edd.), *Biodiversity Conservation and Phylogenetic Systematics: Preserving our evolutionary heritage in an extinction crisis*, Cham, Springer International Publishing, 2016, p. 57–80.

40 J. S. CABRAL – L. VALENTE – F. HARTIG, *Mechanistic simulation models in macroecology and biogeography: state-of-art and prospects*, in «Ecography» 40 (2017), p. 267–280 ; T. F. RANGEL et al., *Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves*, in «Science» 361 (2018), eaar5452.

41 Voir récemment, par exemple, F. BIENVENU – F. DÉBARRE – A. LAMBERT, op. cit.

42 J.-B. Camps – J. Randon-Furling, *Lost Manuscripts and Extinct Texts: A Dynamic Model of Cultural Transmission*, in *Proceedings of the Computational Humanities Research Conference 2022 Antwerp, Belgium, December 12-14, 2022*, Anvers, 2022, p. 198-214 <https://ceur-ws.org/Vol-3290/>.

de pas de temps discrets) et *K* (population maximale). Nous avons fixé l'éventail des valeurs possibles pour ces paramètres sur la base de connaissances historiques ou d'estimations pour la tradition des épopées et des romans de l'ancien français[43]. Nous réalisons, par des simulations, une exploration de l'espace complet de ces paramètres.

Une fois les résultats de ces simulations obtenus, il devient possible de les comparer à des données réelles collectées pour une variété de traditions médiévales[44]. Pour le moment, nos premiers travaux se sont concentrés sur la tradition des chansons de geste et des romans de langue d'oïl. Le principe de cette comparaison est simple : celle-ci se réalise sur la base d'un certain nombre de propriétés observables, tant dans les simulations que les données historiques, telles que la population finale médiane des traditions survivantes (combien de témoin par œuvre), l'âge médian du témoin survivant le plus ancien de chaque tradition ou de l'ensemble des témoins, la proportion de « bifidité » et l'asymétrie des arbres, etc. Si, pour certains paramètres précis, les résultats des simulations concordent avec les données historiques, alors il devient intéressant de consulter les résultats des simulations pour les mêmes paramètres, sur des propriétés non observables dans les données historiques, telles que le taux de survie des œuvres et des témoins, le nombre de générations séparant l'original et l'archétype, l'âge médian des témoins perdus, etc. Le raisonnement est le suivant : si les propriétés observables correspondent aux données historiques, les résultats des modèles pour les mêmes paramètres pourraient fournir des estimations réalistes des propriétés non observables.

Les conclusions actuelles de cette recherche sont les suivantes : un modèle fondé sur le hasard semble permettre d'expliquer une grande partie (mais sans doute pas la totalité) du processus de transmission, survie et extinction des œuvres. Ce modèle produit ainsi des traditions très majoritairement bifides et asymétriques, sans supposer l'existence du biais méthodologique proposé par Bédier.

En outre, pour les chansons de geste et les romans en vers, le modèle permet de proposer des estimations des pertes. Globalement, la valeur la plus haute de survie serait de l'ordre de 55% pour les œuvres et 5% pour les manuscrits, estimations dont on notera avec intérêt qu'elles sont identiques à celles obtenues, pour les mêmes traditions, par Kestemont, Karsdorp *et al.* avec une méthode différente (dite des « espèces non vues »)[45]. Mais il est possible de préciser plus encore en distinguant ces deux corpus : pour les chansons de geste, la survie des œuvres serait de l'ordre de 22 % (1 % pour les manuscrits épiques) et de 33 % pour les romans en vers (5 % pour les manuscrits).

Toutefois, ces recherches n'en sont qu'à leur phase initiale. Il importe ainsi de les prolonger selon différentes approches[46] : élargissement des données envisagées (genre, langue, date des œuvres et manuscrits), mais aussi expérimentation de modèles laissant une part à des déterminismes (par

43 Ces données historiques concernent par exemple la durée moyenne de copie d'un manuscrit, ou bien encore l'éventail de valeur des taux de survie des manuscrits décrits dans des inventaires anciens. Pour plus d'informations, on se reportera à l'article cité.

44 Ces données reposent notamment sur le projet collaboratif et open source OpenStemmata, qui vise à collecter les stemmata publiés par les philologues depuis le XIX[e] siècle ; voir J.B. CAMPS – G. FERNANDEZ RIVA – S. GABAY, Open Stemmata, <https://openstemmata.github.io/>, 2022-… ; ainsi que Eid., *Open Stemmata: A Digital Collection of Textual Genealogies*, in *EADH2021: Interdisciplinary Perspectives on Data, 2nd International Conference of the European Association for Digital Humanities*, Krasnoyarsk, 2021 <https://halshs.archives-ouvertes.fr/halshs-03260086>.

45 M. Kestemont et al., op. cit.

exemple, différents types de sélection), ainsi qu'à la variation dans le temps et l'espace, ou l'expérimentation avec différents taux de changement/d'innovation (traditions plus actives ou plus « quiescentes »). Outre la prise en compte de l'état de l'art des recherches en biologie de l'évolution, ces recherches devront en outre aller plus loin dans la prise en compte des données historiques, en ne regardant pas uniquement des éléments propres à la généalogie ou à la date des témoins, mais aussi en se penchant sur leurs propriétés internes, textuelles notamment.

III. Les méthodes computationnelles, des manuscrits à la critique textuelle

Si le processus de transmission des textes peut être étudié, à un niveau théorique, par des modèles et des simulations, son étude requiert également la prise en compte des données factuelles sur les traditions textuelles survivantes. Là aussi, les méthodes computationnelles, qu'il s'agisse de l'intelligence artificielle ou de l'analyse de données, peuvent être mis à profit, qu'il s'agisse de récolter la matière première des analyses en construisant de vastes corpora textuels, ou bien d'en améliorer la connaissance, en appuyant la critique (d'attribution par exemple). On reviendra donc ici sur l'apport de ces méthodes à l'ecdotique, avant d'évoquer leur contribution à la critique des textes, et on terminera sur certaines perspectives ouvertes par la massification des données.

Ecdotique et chaînes de traitement

La philologie computationnelle s'appuie sur une variété de méthodes nécessitant la puissance de calcul des ordinateurs et des programmes dédiés, allant de l'apprentissage profond à l'analyse de données. Tout cela fait une approche très instrumentée, et donc marquée par un besoin crucial à la fois d'infrastructures et d'équipes comprenant une variété de compétences et de professions : chercheurs, curateurs de données, architectes logiciels et ingénieurs spécialisés dans les chaînes de traitement, sans compter les experts en apprentissage automatique. Grâce à l'Open Source, il est toujours possible pour quelqu'un de procéder seul, avec un investissement non négligeable en compétences de programmation et de science des données, en s'appuyant sur les outils et les modèles largement disponibles. Cependant, en l'absence d'infrastructures, de ressources informatiques et de soutien, on se heurte à des limites pratiques - et aussi un certain gaspillage de temps et de ressources qui pourraient être utilisés à meilleur escient.

Ainsi, un groupe s'est constitué, principalement à l'École des chartes mais avec d'importantes collaborations extérieures, pour travailler depuis plusieurs années à la mise en place de chaînes de traitement et d'infrastructures pour le traitement informatique de documents historiques, notamment de manuscrits médiévaux. Conçu et utilisé dans le cadre de plusieurs doctorats[47], et soutenu par de nombreux projets différents, ce flux de travail modulaire a abouti à la réalisation et à la publication

46 Ces enquêtes futures sont l'objet principal du projet LostMa: The Lost Manuscripts of Medieval Europe: Modelling the Transmission of Texts (ERC StG, 2024-2028), que je dirige, et qui prolongera et étendra cette méthodologie aux traditions européennes médiévales des textes épiques et romanesques.

47 J.-B. CAMPS, *La `Chanson d'Otinel': édition complète du corpus manuscrit et prolégomènes à l'édition critique*, thèse de doctorat, dir. D. BOUTET, Paris, Université Paris-Sorbonne, 2016 ; A. PINCHE, *Édition nativement numérique du recueil hagiographique «Li Seint Confessor» de Wauchier de Denain d'après le manuscrit fr. 412 de la Bibliothèque nationale de France*, thèse de doctorat, dir. C. PIERREVILLE et B. BUREAU, Lyon, Université Lyon 3, 2021 ; T. CLÉRICE, *Détection d'isotopies par apprentissage profond : l'exemple de la sexualité en latin classique et tardif*, thèse de doctorat, dir. C. NICOLAS, Lyon, Université Lyon 3, 2022 ; L. ING, *L'obsolescence lexicale en français médiéval. Philologie et linguistique computationnelles sur le Lancelot en prose*, thèse de doctorat, dir. F. DUVAL et codir. J.B. CAMPS, Paris, École nationale des chartes, 2023.

de logiciels, de données, de modèles et de formats pour l'analyse de la mise en page[48], la transcription automatique des manuscrits (*Handwritten Text Recognition*, ou HTR) médiévaux et d'imprimés anciens[49], incluant notamment les problématique de la resegmentation en mots linguistiques modernes ou du traitement des abréviations[50], l'annotation linguistique du latin, du français médiéval et moderne[51] ou de la collation[52].

Cette chaîne de traitement se veut modulaire : chacune de ses étapes (analyse de la mise en page, transcription automatique, normalisation, annotation linguistique, alignement et collation) est optionnelle, peut donner lieu à des corrections manuelles ou non (selon l'objectif recherché), et peut se voir substituer un nouvel outil si besoin. Ce dernier point est crucial, car, d'une part, les méthodes de traitement automatique des langues évoluent à un rythme de plus en plus rapide, et que, d'autre part, les données librement disponibles et la qualité des modèles qui en résultent amènent à des mises à jour pluri-annuelles[53].

Plus encore que la performance des algorithmes, c'est la libre disponibilité de données vérifiées et de qualité qui constitue le facteur limitant des chaînes de traitement. Pour cette raison, outre une politique d'*Open Source* revendiquée, un soin particulier a été porté aux choix ou au développement d'outils ergonomiques facilitant ces tâches intensives en temps et en travail humain expert[54].

Critique des textes et analyse de données

L'application des méthodes computationnelles ne se limite pas aux questions d'édition des textes, mais touche également à leur critique, qu'il s'agisse, d'une part, de critique d'attribution, datation ou localisation, et d'autre part, de critique des variantes. Ces deux types de critique trouvent leur application tant à l'histoire d'un texte donné, qu'à l'histoire culturelle au sens large, car elles peuvent être révélatrices des facteurs déterminants individuels et collectifs du style, ainsi que du processus de variation textuelle en lui-même. En tant que telles, elles peuvent venir soutenir l'étude des grands mécanismes de transmission et d'évolution des textes.

---

48 S. GABAY et al., *SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more)*, in *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*, 2021.

49 A. PINCHE, *CREMMA Medieval, an Old French dataset for HTR and segmentation*, v1.0 (2021), <https://github.com/HTR-United/cremma-medieval> ; C. JAHAN – S. GABAY, *OCR17+ - Layout analysis and text recognition for 17th c. French prints*, v1.0 (2021), <https://github.com/e-ditiones/OCR17plus> ; T. CLÉRICE – A. PINCHE – M. VLACHOU-EFSTATHIOU, *Generic CREMMA Model for Medieval Manuscripts (Latin and Old French), 8-15th century*, in «Zenodo», 2023, doi:10.5281/zenodo.7631619.

50 J.-B. CAMPS – C. VIDAL-GORÈNE – M. VERNET, *Handling Heavily Abbreviated Manuscripts: HTR engines vs text normalisation approaches*, in *International Conference on Document Analysis and Recognition*, Springer, 2021 ; J.-B. CAMPS et al., *Data Diversity in handwritten text recognition: challenge or opportunity?*, in *Digital Humanities Abstract, DH2022 - Tokyo*, Tokyo, 2022, ainsi que T. CLÉRICE, *Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin*, in «J. Data Min. & Digit. Humanit.» 2020, doi:10.46298/jdmdh.5581.

51 J.-B. CAMPS et al., *Corpus and Models for Lemmatisation and POS-tagging of Old French*, in «ArXiv» (2021), 210911442 Cs ; J.-B. CAMPS et al., *Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre*, in «J. Data Min. & Digit. Humanit.» 2021, doi:10.46298/jdmdh.6485.

52 J.-B. CAMPS – L. ING – E. SPADINI, *Collating Medieval Vernacular Texts: Aligning Witnesses, Classifying Variants*, in *Digital Humanities Conference 2019, Complexities*, Utrecht, 2019.

53 Pour un aperçu complet de cette chaîne de traitement, désormais quelque peu daté, on pourra se reporter à J.-B. CAMPS, *La philologie computationnelle à l'École des chartes*, cit.

54 Pour l'HTR, le choix s'est porté sur la plateforme eScriptorium (B. KIESSLING, i, in *Digital Humanities Conference 2019, Complexities*, Utrecht, 2019), tandis que, pour la correction de l'annotation linguistique, le logiciel Pyrrha a été développé (T. CLÉRICE et al., Pyrrha, *A language independant post correction app for POS and lemmatization*, v4.0 [2024], doi:10.5281/zenodo.2325427).

La critique d'attribution est un des champs d'application principaux de la stylométrie, la mesure du style, ou, pour être plus précis, de l'idiolecte des auteurs[55]. Cette discipline, qui remonte au XIX^e^ siècle[56], a été profondémment renforcée par la puissance combinée des ordinateurs et du développement des méthodes statistiques pendant la deuxième moitié du XX^e^ siècle. Elle s'attache généralement aux propriétés de langue les moins consciemment utilisées et les plus détachées des variations volontaires et conscientes (génériques notamment). Pour cette raison, elle se concentre souvent sur l'emploi des mots-outils et morphèmes grammaticaux[57]. Mais la variation idiolectale n'est jamais parfaitement isolée, et la stylométrie doit ainsi accepter l'existence de déterminants sociaux et culturels collectifs, tels que le genre ou l'âge des individus. D'autres contraintes collectives jouent un rôle central, comme notamment les genres littéraires, ou les conditions matérielles de production et de transmission des textes. La stylométrie suit ainsi une direction parallèle à celle de Barthes, qui proposait de « dépasser la notion d'idiolecte [primitivement retenue comme point de départ] et à voir dans toute écriture, fût-elle apparemment très individuelle, le fragment d'un sociolecte ou langage de groupe »[58].

Pour le spécialiste des textes médiévaux de langue d'oïl, la stylométrie a une pertinence particulière, due à l'importance de l'anonymat ou des attributions contradictoires dans la tradition des textes : par exemple, 95 % environ des chansons de geste sont anonymes, tandis que, sur les 2862 chansons de trouvères connues, 65% sont anonymes, 17% ont au moins deux attributions dans la tradition manuscrite, et seulement 20% sont dotées d'une attribution unique (ce qui ne signifie pas nécessaire une attribution fiable)[59]. Et pourtant, leur analyse pose des difficultés particulières : en particulier, les scribes ont tendance à apporter de nombreuses modifications aux mots-outils et morphèmes grammaticaux qui constituent une des matières premières principales de la stylométrie. Cette difficulté a parfois conduit à rechercher des traits plus stables, comme les choix de rimes, qui ne sont pratiques que pour des textes relativement longs[60]. Plus répandues encore que les substitutions de mots-outils sont les variantes graphiques, qui peuvent nécessiter une normalisation ou une lemmatisation avant analyse[61].

Sur ces bases, les questions qui se posent à la stylométrie des œuvres médiévales sont les suivantes : Peut-on encore reconnaître les auteurs des textes médiévaux conservés dans des

55 S. GABAY, *Beyond Idiolectometry? On Racine's Stylometric Signature*, in M. EHRMANN et al. (Edd.), *Proceedings of the Conference on Computational Humanities Research 2021*, Amsterdam, 2021 (CEUR Worshop Proceedings, 2989), p. 359-376 <https://ceur-ws.org/Vol-2989/>.

56 W. LUTOSLAWSKI, *Principes de stylométrie appliqués à la chronologie des œuvres de Platon*, in «Rev. Études Grecques» 11 (1898), p. 61–81 ; T. C. MENDENHALL, *The characteristic curves of composition*, in «Science» (1887), p. 237–246.

57 Et ce depuis W. DITTENBERGER, *Sprachliche Kriterien für die Chronologie der Platonischen Dialoge*, in «Hermes» 16 (1881), p. 321–345. Voir plus récemment M. KESTEMONT, *Function words in authorship attribution : From black magic to theory?*, in *Proceedings of the 3rd Workshop on Computational Linguistics for Literature* (CLFL), 2014, p. 59-66 <https://aclanthology.org/W14-0908.pdf>.

58 R. BARTHES, *Œuvres complètes, vol. 3 (1968-1971)*, éd. par É. MARTY, Paris, Éditions du Seuil, 2002., cité par S. Gabay, *Beyond Idiolectometry…*, cit.

59 D'après les données collectées par L. GATTI, *Repertorio delle attribuzioni discordanti nella lirica trovierica*, Roma, Sapienza Università Editrice, 2019.

60 M. KESTEMONT – W. DAELEMANS – D. SANDRA, *Robust Rhymes? The Stability of Authorial Style in Medieval Narratives*, in «Journal of Quantitative Linguistics» 19 (2012), p. 54–76.

61 M. KESTEMONT – S. MOENS – J. DEPLOIGE, *Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux*, in «Lit. Linguist. Comput.» 30 (2013 2015), p. 199–224 ; J.-B. CAMPS – T. CLÉRICE – A. PINCHE, *Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer's hagiographic hypothesis*, in «Digit. Scholarsh. Humanit.» 36 (2021), p. ii49–ii71; J.-B. CAMPS – B. SALVATI, *On Burgundian (di)vine orators and other impostors: Stylometry of Late Medieval*

manuscrits ? Le signal de l'idiolecte de l'auteur survit-il au processus de transmission ? Peut-il aider à déterminer les sources des grandes compilations d'œuvres antérieures ?

Sur ces points, nous avons récemment réalisé une étude portant sur le cas des légendiers français en prose. L'histoire de la constitution de ces compilations de vies de saints reste en partie une énigme, car les manuscrits conservés montrent des compilations déjà constituées, mais il nous manque les premières étapes qui restent sujettes à conjectures. Ce cas n'est pas isolé pour les textes médiévaux, et présente des similitudes avec la situation des anthologies poétiques des troubadours et des trouvères. Les hypothèses concernant ces compilations postulent que leur organisation est tributaire de celle de leurs sources, qui auraient pu être des livres antérieurs plus petits (perdus), généralement appelés *libelli*. Ces *libelli* auraient pu être des unités cohérentes en soi, des regroupements de textes partageant un auteur, un compilateur ou un thème commun (par exemple, un *libellus* sur un groupe spécifique de saints). Paul Meyer a fameusement étudié la macrostructure des légendiers et fait l'hypothèse que les manuscrits que nous conservons sont le résultat de plusieurs phases successives de compilation, distinguant plusieurs collections (auxquelles il attribue des sigles différents)[62]. Il émet également l'hypothèse que ces collections dépendent de séries plus petites préexistantes, dont certaines sont identifiables par leur auteur, comme les vies des Saints Confesseurs de Wauchier de Denain, bien que la plupart des vies des compilations restent anonymes. Une analyse stylométrique (appuyée sur une chaîne de traitement ayant permis l'acquisition, segmentation, et lemmatisation du texte du ms. BnF, fr. 412) ont permis de confirmer la cohérence stylistique des œuvres attribuées à Wauchier de Denain, qui se regroupent dans toutes les analyses en un ensemble très cohérent. Ces analyses ont également fourni des résultats assez similaires aux hypothèses de Paul Meyer, avec quelques notables exceptions qui permettent de raffiner le classement qu'il avait proposé. Par exemple, la vie de saint Longin a été placée par Meyer dans la collection A, probablement parce qu'il a utilisé comme base le seul manuscrit où c'est le cas (BnF, nouv acq. fr. 23686) ; notre analyse la place en revanche dans la collection B, comme dans la plupart des manuscrits survivants. Par ailleurs, nous proposons d'identifier plusieurs sous-séries précises[63].

Outre la stylométrie, la scriptométrie (qui diffère plutôt par son objet que par ses méthodes) est une approche pertinente pour soutenir la datation et la localisation des œuvres. Inaugurée pour le domaine d'oïl par Dees, et poursuivie par H. Goebl, elle peut désormais bénéficier de la croissance des matériaux disponibles[64].

L'étude computationnelle des variantes est un domaine encore embryonnaire. Pourtant, depuis Havet et, pour le domaine d'oïl, Robert Marichal, l'idée est présente de construire des typologies de

---

Rhetoricians, in *DH 2023: Collaboration as Opportunity*, Graz, 2023, p. 159-163, doi:10.5281/zenodo.8210808.

62 P. MEYER, *Légendes hagiographiques en français*, «Histoire littéraire de la France», 33 (1906), p. 328–458.

63 CAMPS – CLÉRICE – PINCHE, *Noisy medieval data*…, cit.

64 A. DEES – P. VAN REENEN – J. A. DE VRIES, *Atlas des formes et des constructions des chartes françaises du XIIIe siècle*, Tübingen, M. Niemeyer Verlag, 1980 ; DEES et al., *Atlas des formes linguistiques des textes littéraires de l'ancien français*, Tübingen, M. Niemeyer, 1987 ; H. GOEBL, *L'aménagement scripturaire du Domaine d'Oïl médiéval à la lumière des calculs de localisation d'Anthonij Dees effectués en 1983: une étude d'inspiration scriptométrique*, in «MR», *Seminario 2011: Il problema della scripta, Venezia, 13-14 ottobre 2011* (2011); J.-B. CAMPS, *Manuscripts in Time and Space: Experiments in Scriptometrics on an Old French Corpus*, in A. U. FRANK et al., *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)*, Vienne, 2018, p. 55–64.

variante et les mettre en série[65]. De rares travaux récents se sont penchés sur des traditions artificielles, et ont montré que la correspondance entre évaluation par les experts et valeur généalogique réelle des variantes restait problématique[66]. Pour progresser dans ce domaine, trois éléments principaux sont nécessaires : une collecte significative de variantes, des typologies permettant de classer les données recueillies, et des analyses portant à la fois sur des traditions historiques réelles et d'autres produites dans des contextes experimentaux contrôlés, comme cela se fait en sciences cognitives[67].

Perspectives de la massification des données

La massification des données permise par l'intelligence artificielle ouvre de nombreuses perspectives pour articuler les différents niveaux d'analyse, de la lecture proche des textes aux études macrostructurelles de longue durée, et pour lier réflexion théorique et modélisatrice avec prise en compte des données textuelles.

Un exemple de problématique possible est ainsi le lien éventuel entre l'importance accordée aux contenus consacrés à l'amour courtois et le volume de la production (et survie) de copies des textes concernés. Dans une perspective encore exploratoire, nous avons construit à cette fin un corpus de manuscrits des fictions narratives longues de langue d'oïl, du XIII^{e} au XV^{e} siècle. Les numérisations de 409 témoins ont ainsi été récupérées sur la plateforme Gallica de la Bibliothèque nationale de France, et traitées par transcription automatique (HTR) pour constituer un corpus de 40 millions de mots, répartis sur trois siècles, le *Corpus of Medieval French Epics and Romances*[68]. Si la répartition chronologique de son contenu est tributaire des aléas de la production et de la conservation des manuscrits, et laisse à voir des périodes creuses liées à la Peste Noire et à la guerre civile entre Armagnacs et Bourguignons, l'évolution dans le temps de l'importance des motifs narratifs amoureux dans les textes semble présenter une phase d'expansion au XIII^{e} et au début du XIV^{e} siècle, suivie d'une période de contraction qui semble correspondre à la crise du Moyen Âge tardif, ce qui n'est pas sans évoquer l'hypothèse de Georges Duby, qui liait floraison de l'amour courtois et phases de développement économique[69].

*****

******

Plus qu'une simple poursuite des questionnaires philologiques actuels par d'autres moyens, la philologie computationnelle peut, dans certains contextes, amener des déplacements ou des transformations dans les paradigmes de recherche en sciences des textes. Sans sortir du paradigme

65 L. HAVET, *Manuel de critique verbale appliquée aux textes latins*, Paris, Librairie Hachette, 1911 ; R. MARICHAL, *Conclusion*, in *La pratique des ordinateurs dans la critique des textes*, Paris, CNRS Éditions, 1979, p. 285-288.

66 T. L. ANDREWS, *Analysis of variation significance in artificial traditions using Stemmaweb*, in «Digit. Scholarsh. Humanit.» 31 (2016), p. 523–539.

67 Depuis 2022, un séminaire établi à l'École des chartes, dirigé par Frédéric Duval, Benedetta Salvati et moi-même, s'est fixé la tâche de produire une typologie nécessaire à la description des variantes, selon leur nature, leur genèse et leur évaluation critique. Ce type de grille peut être mis à profit pour, par exemple, évaluer sur des traditions le niveau de cohérence entre types de variantes et hypothèses stemmatiques (B. SALVATI – J.-B. CAMPS, *Combining the Encoding of Variants with Stemmatological Analysis: the case of Chrétien's 'Cligés'*, article accepté pour la conférence *Studia Stemmatologica Conference 2024, Monte Verità*).

68 CAMPS et al., *Make Love or War?*…, cit. Pour le corpus actuel, voir CAMPS, *Corpus of Medieval French Epics and Romances*, v1.0 (2023), doi:10.5281/zenodo.8208122.

69 G. DUBY, *Mâle Moyen Âge: de l'amour et autres essais*, Paris, Flammarion, 1987 ; voir aussi N. BAUMARD et al., *The cultural evolution of love in literary history*, in «Nat. Hum. Behav.» 6 (2022), p. 506–522.

évolutionniste qui remonte aux origines de la discipline, la philologie peut ainsi reprendre à son compte des approches méthodologiques articulant démarche de test d'hypothèse, modèles et analyse de données. La massification des données permise par l'intelligence artificielle ouvre des perspectives d'analyse nouvelles, propice aux études macrostructurelles, comparatistes ou de longue durée.