Ancien ou moderne ? Pistes computationnelles pour l'analyse graphématique des textes écrits au XVIIe siècle

Simon Gabay, Philippe Gambette, Rachel Bawden, Benoît Sagot

Ancien ou moderne ? Pistes computationnelles pour l'analyse graphématique des textes écrits au XVII^{e} siècle

Old or Modern? Towards a Computational Graphematic Analysis of 17th Century French Texts

Simon Gabay, Philippe Gambette, Rachel Bawden et Benoît Sagot

1. Introduction

Les systèmes graphiques du français du XVII^{e} s. restent peu étudiés, en dépit de leur grande richesse. Contrairement à d'autres époques ou d'autres phénomènes (syntaxiques, lexicaux, etc.), cette sous-exploitation concerne aussi la variété la plus prestigieuse de la langue, celle des auteurs dits classiques. Les éditions de leurs textes restent en effet encore douteuses (Duval, 2015), notamment concernant le vêtement graphique original qui est (presque) toujours abandonné au profit du français contemporain (Gabay, 2014), rendant l'analyse graphématique de la langue du Grand Siècle plus laborieuse que pour les époques précédentes.

La linguistique descriptive est tout particulièrement concernée par cet oubli, d'autant plus problématique que le français, langue de Molière, ne cesse de désigner le XVII^{e} s. comme sa principale source, notamment pour l'orthographe. La persistance de cet angle mort n'est pas souhaitable pour de multiples raisons, comme celle d'une perception incomplète du changement linguistique en français – question que nous avons récemment abordée ailleurs (Gabay *et al.*, 2022) – mais pas uniquement. D'autres pans de la romanistique sont aussi concernés, comme l'ecdotique. Les nouvelles possibilités offertes par l'informatique permettent désormais l'automatisation de certaines tâches, comme les relevés et l'analyse linguistiques, qui s'intéressent de près

aux graphies : il nous faut donc fournir aux éditeurs les moyens non seulement intellectuels, mais aussi techniques de simplifier et accélérer l'étude des systèmes graphiques auxquels ils se retrouvent confrontés, et ainsi favoriser l'établissement de textes de meilleure qualité, présentant une langue dont le vêtement graphique est moins retouchée et mieux décrite.

Le rendement interprétatif de l'étude des systèmes graphiques dépasse cependant de loin ce simple cadre linguistico-ecdotique. Recoupant en grande partie l'opposition entre les Anciens et les Modernes (Biedermann-Pasques, 1992), les tergiversations graphématiques des auteurs rendent potentiellement compte de prises de positions politiques, informant de ce fait, par des voies inhabituelles, l'histoire de la littérature. Le cas de Pierre Corneille est certainement le plus célèbre (Pellat, 1992), mais n'est pas unique : si les imprimeurs ont joué un rôle important dès le XVI^e^ s. (Baddeley, 1996 et Riffaud, 2007), l'étude attentive des imprimés et des manuscrits permet de retrouver la langue des auteurs, et ainsi de révéler des trajectoires intellectuelles, des dissonances idéologiques ou des logiques de circulation textuelle.

Après un bref retour sur la question des grands courants et écoles orthographiques du français, avec des systèmes graphiques recatégorisés pour l'occasion en *scriptae*, nous nous pencherons sur la conception d'une typologie propre à l'analyse scriptométrique. Si la détection automatique des *scriptae* doit en effet s'appuyer sur les travaux des linguistes, la logique computationnelle n'est pas strictement identique aux approches traditionnelles et nécessite des aménagements pour des raisons techniques comme méthodologiques. Cette approche étant nouvelle en linguistique diachronique, nous terminons cet article en l'évaluant sur un cas connu (celui des imprimés de Boileau) pour démontrer l'intérêt d'une analyse computationnelle des *scriptae* à l'époque classique, à même d'être reproduite par tous les éditeurs et linguistes qui pourraient être intéressés.

2. Des *scriptae* en français classique

Si la littérature scientifique tend à parler d'« orthographe » pour décrire la forme écrite du français des XVI^e^ et XVII^e^ s., le terme nous parait problématique, car aucune « forme correcte » ne peut à cette époque être déduite d'un standard toujours en formation. Plutôt qu'une orthographe, il nous semble qu'on observe des « usages individuels » et des « systèmes graphiques », pour reprendre la terminologie des médiévistes[1]. Ces systèmes graphiques, qu'il faut distinguer de leur réalisation orale, sont constitués de faisceaux de traits cohérents, restent perméables les uns aux autres et peuvent être rassemblés en grands groupes, à la manière des *scriptae* médiévales – un terme que nous nous proposons de transposer à l'époque moderne.

Les médiévistes pourront s'étonner de cette utilisation du terme, mais rappelons que Remacle lui-même précise que « l'expression 'la scripta' est synonyme de l'allemand 'die Schriftsprache' » (1948 : 21) : la *scripta* est donc tout simplement une langue écrite. L'objectif de Remacle de distinguer « un trait du dialecte réel [de] ce qui est seulement un caractère graphique de la scripta » (Remacle, 1948 : 27) n'est alors qu'un des usages possibles d'une *Schrifsprache*. Remacle précise aussi que la *scripta* désigne « la langue vulgaire écrite au moyen âge » (1948 : 21), ce qui fait sens étant donné l'importance des traits dialectaux présents dans la langue écrite médiévale, mais pourrait être perçu comme trop restrictif si l'on pense à la persistance de tels traits aux siècles suivants

(Bergeron-Maguire, 2019). Par ailleurs, si « [Skripta] bedeutet zunächst 'mittelaterliche Schriftsprache' und umgeht im Französischen die Vieldeutigkeit der Bezeichnungen 'langue écrite' und 'langue littéraire' » comme l'explique Gossen (1967 : 5), pourquoi encore une fois restreindre l'usage de *scripta* au Moyen Âge ? L'ambiguïté des termes « langue écrite » et « langue littéraire » ne nous semble pas propre au Moyen Âge.

Nous rejoignons la définition de St. Koch, selon qui le terme de *scripta* a été conçu « para designar textos no latinos en la 'lengua vulgar' (romance) antes de la existencia de un estándar en el sentido moderno » (Koch, 2013 : 596)[2]. Il ne s'agit nullement de nier la spécificité de la langue médiévale, sur laquelle insistent tous les chercheurs que nous venons de citer, mais l'étude des *Schriftsprachen* avant la standardisation ne peut se limiter uniquement à l'étude de la correspondance entre le code graphique et une variation phonétique : il convient d'ajouter à la dimension diatopique de la *scripta* d'autres dimensions[3], comme la dimension diastratique, afin de rendre compte de sociolectes actifs pendant le Grand Siècle.

Tableau 1. Exemples d'opposition entre graphie ancienne et graphie moderne

	Anciens	**Modernes**
Lettres ramistes	Position (*vniuers*)	Dissimilation (*univers*)
Ancien hiatus	maintenu (*veu*)	supprimé (*vu*) ou accent circonflexe (*vû*)
Lettre calligraphique	maintenue (*ay*)	supprimée (*ai*)
Pluriel nominal	*-s*, *-z*, *-x*	*-s*
Consonne muette (diac., étym., hist.)	maintenue (*doubte*)	supprimée (*doute*)
Voyelles longues	dédoublement (*aage*)	circonflexe (*âge*)

La critique a retenu une tension entre deux grandes tendances à l'époque moderne qui, si elles rendent imparfaitement compte de toute la richesse de la variation, permettent une compréhension des grands enjeux politico-linguistiques de l'époque. Nous trouvons d'une part l'« orthographe ancienne », qui maintient contre l'évolution de la langue le lien avec substrat latin, et d'autre part l'« orthographe moderne », qui milite pour l'adoption d'un code graphique mieux adapté au français. En pratique, cette opposition théorique s'organise notamment (mais pas uniquement) en trois grands pôles (cf. tab. 1)[4] :

> le système alphabétique, avec les couples <i>/<j> et <u>/<v> (par ex. *vniuers* vs *univers*) conçus comme des variantes graphiques (majuscule, initiale...) ou phonogrammiques (phonème voyelle vs. phonème consonne) ;
> les logogrammes lexicaux, avec l'utilisation ou la suppression de consonnes muettes, qu'elles soient historiques (*donner* vs *doner* < lat. *donare*) ou étymologiques (*douter* vs *doubter* < lat. *dubitare*) ;
> les signes auxiliaires, avec l'utilisation de lettres diacritiques plutôt que d'accents ou de trémas (*hôpital* vs *hospital* < empr. au lat. *hospitalis*).

Plus que deux « orthographes », ce système d'opposition renvoie à des courants profonds, très précocement actifs dans l'histoire du français avec d'une part l'adoption d'un code graphique « correspondant à la tradition romane primitive » (Pellat, 1994), et d'autre part une volonté de relatiniser le français qui apparait dès le XIV^e^ s. (Brazeau et Lusignan, 2004). Il convient donc de différencier ces grands courants de leur réalisation à l'époque moderne, raison pour laquelle nous souhaitons réintroduire le terme de *scripta* afin de décrire un type de variation linguistique propre à cette époque, en prenant soin de distinguer les *scriptae* romanes, nées spontanément du latin, des *scriptae* françaises, apparues en français après le Moyen Âge.

Nous nous proposons de distinguer deux *scriptae* françaises : celle « des Anciens », défendue par quelques grammairiens de l'époque voulant inscrire l'histoire du français dans sa forme écrite, et celle « des Modernes », rompant avec cet héritage linguistique au profit d'innovations pensées comme salutaires. Cette terminologie choisit délibérément de s'aligner sur les deux factions animant la célèbre Querelle des Anciens et des Modernes (Fumaroli, 2001), dont elle rejoue linguistiquement l'opposition intellectuelle entre tradition et réforme, vis-à-vis de laquelle chacun était tenu à l'époque de se positionner, comme le fit Bossuet (1627-1704) dans les *Cahiers de Mezeray* :

> [La compaignie] ne peut souffrir une fausse regle qu'on a uoulu introduire, d'écrire comme on prononce, parce qu'en uoulant instruire les étrangers et leur faciliter la prononciation de nostre langue, on la fait meconnoistre aux François mesmes. Si on ecrivoit *tans*, *chan*, *cham*, *emais* ou *émés*, *connoissans*, *anterreman*, *faisaict*, qui reconnoistroit ces mots ? [...] Il y a aussi une autre ortographe, qui s'attache scrupuleusement a toutes les lettres tirées des langues dont la nostre a pris ses mots, et qui ueut ecrire *nuict*, *ecripture*, etc. Cella blesse les yeux d'une autre sorte en leur remettant en ueüe des lettres dont ils sont desaccoustumez et que l'oreille n'a iamais connus [*sic*]. (*Cahiers de remarques...*, 1863 : xiv).

Un coup d'œil sur la documentation manuscrite révèle, derrière ces déclarations théoriques pourtant limpides de 1673, toute la complexité de la pratique personnelle de Bossuet. L'abbé Lebarq, qui s'est longuement penché sur les autographes lors de l'édition des *Œuvres oratoires* de l'Aigle de Meaux, a ainsi démontré la variation du système graphique bossuetien au cours du temps et les potentialités herméneutiques de cette évolution (Bossuet 1890, t. 2 : vi ; cf. tab. 2).

Tableau 2. Exemples de l'évolution du système graphique de Bossuet (le mot-vedette est la version contemporaine des formes classiques).

Manuscrit	**tant**	**être**	**cette**	**même**	**paraître**	**avec**	**a-t-il**
BNF Fr. 12822, f°370, Brièveté de la vie, 1648 (Bossuet, 1890, t. 2 : vi).	tans	estre	Ceste	mesme	parêtre	auecque	a t'il
BNF Fr. 12823, f°130, Fête du Rosaire, oct. 1651 (Bossuet, 1890, t. 2 : 61).	tans	être	cête	même	parêtra	auec	a t'il
BNF Fr. 12824, f°119, Sur la Providence, mai 1656 (Bossuet, 1890, t. 2 : 146)	temps	estre	ceste	mesme	paroist	auecque	a til
BNF Fr. 12822, f°61, Sur les démons, fév. 1660 (Bossuet, 1890, t. 2 : 213)	temps	estre	cette	mesme	Paroistre	-	atil

Les deux premiers manuscrits, qui ont été écrits avant la fin de ses études (reçu docteur en théologie en 1652), sont marqués par une *scripta* des Modernes qui voit la suppression des lettres historiques et diacritiques (*meme*), parfois au profit d'accents (ici circonflexe). Les deux autres manuscrits, écrits alors que Bossuet a rejoint les ordres en Lorraine (1654), témoignent d'un retour rapide à une *scripta* des Anciens et l'abandon des marqueurs typiques de la modernité à quelques exceptions près (*cette*). Par l'étude du système graphique, c'est un parcours idéologique qui se dessine, marqué par un lent repli sur la tradition après les années d'université.

Aux revirements idéologiques, et parfois plus simplement aux effets du temps, il convient d'ajouter l'inconsistance personnelle comme source de variation graphématique[5], ainsi que l'observaient déjà des contemporains comme Cl. Buffier :

> Il faut observer que ces deux sortes d'ortographe étant en usage, il arrive non-seulement qu'elles sont employées, l'une par certains Auteurs, & l'autre par d'autres ; mais aussi que le même Auteur prenant quelquesfois l'une pour l'autre, les emploie toutes deux sans y penser en divers endroits de ses ouvrages, ou même qu'il suit l'une en certains chefs & l'autre en d'autres chefs. (Buffier 1709 : 410).

Le phénomène décrit ici peut certainement être rapproché de la notion d'« orthographe actuelle », utilisée par Cl. Vachon pour décrire une *Sprachmischung* (« mélange des langues ») retenant par exemple des traits des Anciens dans une *scripta* des Modernes. (Vachon, 2010 : 251).

3. Typologies pour l'analyse des *scriptae* du XVII^e^ s.

L'étude de la variation graphique au XVII^e^ s. nécessite une typologie afin de regrouper les occurrences et d'en faire sens. Si plusieurs chercheurs ont proposé des classements, ces derniers posent restent complexe à opérationnaliser d'un point de vue informatique, car ils s'appuient sur des distinctions que la machine n'est pas (encore) capable de faire. L'utilisation d'ordinateurs permet cependant des relevés dans de très grands ensembles, dont il ne faut pas sous-estimer l'utilité.

Comme nous l'avons dit en introduction, l'étude des systèmes graphiques après le Moyen Âge et avant la standardisation (au moins théorique) du français reste pour partie un angle mort de la recherche. Les travaux fondateurs de Nina Catach (2001) se concentrent essentiellement, pour la période classique, sur les propositions des Remarqueurs et celles de l'Académie, sans étudier dans le détail le contenu des textes qui nous sont parvenus. De telles études ont été menées plus tardivement par Biedermann-Pasques (1992), Pellat (1994) et Vachon (2010), qui reprennent les catégories d'Ancien et de Moderne. Les deux derniers ont tenté de proposer des typologies claires, qui divergent l'une de l'autre. Pellat propose une brève liste de quatorze variantes graphiques majeures, pour lesquelles il oppose à chaque fois la version ancienne et la version moderne :

1. Distinction <i>/<j> et <u>/<v>. Par ex. : *ie* vs *je* ou *vniuers* vs *univers* ;
2. Graphies de [s] (<c>, <ç>, <s>, <ss>, <sc>). Par ex. : *sçavoir* vs *savoir* ou *receu* vs *reçu* ;
3. Notation des voyelles longues. Par ex. : *aage* vs *âge* ou *preste* vs *prête* ;
4. Timbre de E :
    a. fermé. Par ex. : *estoit* vs *était* ;
    b. moyen. Par ex. : *fidelle* vs *fidèle* ;
    c. ouvert (bref). Par ex. : *apres* vs *après* ;

5. Lettre <y> :

a. semi-consonne. Par ex. : *ayeux* vs *aïeux* ;
b. calligraphique (par ex. : *soy* vs *soi*) ou diacritique (par ex. : *suyure* vs *suivre*) ;
c. hiatus. Par ex. : *pays* vs *païs* ;

6. Digraphe vocalique :

<au> (par ex. : *pseaume* vs *psaume*) et <oi> (par ex. : *connoistre* vs *connaître*) ;
ancien hiatus. Par ex. : *deu* vs *dû* ;

7. Accent grave : <à> (par ex. : *a* vs *à*) et <ù> (par ex. : *ou* vs *où*) ; tréma (par ex. : *poete* vs *poëte*) ;
8. Voyelles nasales. Par ex. : *avanture* vs *aventure* ;
9. Consonnes doubles (diacritiques et historiques). Par ex. : *addresser* vs *adresser* ;
10. Consonnes muettes (<c>, <d>, <l>, <g>, <p>, <cq>) étymologiques (par ex. : *soubdain*<lat. *subitānus* vs *soudain*), historique (par ex. : *commung* vs *commun*), diacritique (par ex. : *beautez* vs *beautés*) ;
11. Lettres grecques. Par ex. : *cholere* (<lat. *cholera*, du grec χολέρα> vs *colère* ;
12. Finales plurielles :

a. pluriels nominaux. Par ex. : *loix* vs *lois* ;
b. finales de [e]. Par ex. : *venés* vs *venez* ;
c. finale <ant>/<ent>+<s>. Par ex. : *perdans* vs *perdants* ;

13. Conjugaisons verbales :

a. P1 du présent de l'indicatif. Par ex. : *je voy* vs *je vois* ;
b. P2 de l'impératif. Par ex. : *di* vs *dis* ;

14. Accord des participes :

a. participe présent vs adjectif verbal (invariable ou non) ;
b. participe passé (accord avec le participe passé antéposé avec avoir) ;

À cette description mettant en avant les principaux lieux variants s'ajoute celle de Cl. Vachon, qui concerne, rappelons-le, le XVI^e^ s. Contrairement à J.-Ch. Pellat, elle n'articule pas son propos autour de l'opposition graphie ancienne / graphie moderne, mais autour de l'opposition forme concurrente / forme actuelle (la première forme ayant pour caractéristique d'avoir été abandonnée, cf. Vachon, 2010 : 73). Il faut donc déduire, pour chaque cas, ce qui relève des Modernes et des Anciens. Les différentes formes sont classées (plutôt) phonétiquement[6] :

1. Voyelles :

a. Orales. Par ex. : *oblier* vs. *oublier* ;
b. Diphtongues (par ex. : *affoiblir* vs. *affaiblir*) et hiatus (par ex. : *accreu* vs. *accru*) ;
c. Nasales. Par ex. : *menger* vs. *manger* ;

2. Consonnes :

a. Doubles. Par ex. : *atendre* vs. *attendre* ;
b. Muettes et internes. Par ex. : *effect* vs. *effet* ;
c. Distinction <i>/<j> et <u>/<v>. Par ex. : *ie* vs *je* ou *vniuers* vs *univers* ;

3. Variations purement morphographiques :

a. Régularisation du pluriel des thèmes en <-t> : <z>/<ts> et <s>/<ts>. Par ex. : *petiz* vs. *petits* ;

b. Morphème adverbial <s>. Par ex. : *encores* vs. *encore* ;
c. Désinences verbales. Par ex. : *je rend* vs. *je rends* ;

On le voit, ces deux classements convergent amplement dans leurs constats sur les lieux et les types de variation à de rares exceptions près – ainsi la question des graphies du phonème [s] n'est qu'imparfaitement traitée par Cl. Vachon, qui ne mentionne pas l'alternance <s>/<c> (*défense* vs *défence*), mais J.-Ch. Pellat ne mentionne pas le cas du morphème adverbial <-s> (*encores*). Une typologie plus précise des lieux des variations reste donc encore à établir, mais tel ne peut être notre objet ici pour plusieurs raisons.

Premièrement, il est impossible d'attendre de la machine qu'elle puisse opérer certains classements à l'heure actuelle. Ainsi, en l'absence d'informations étymologiques computationnellement exploitables, il reste impossible de distinguer parmi les consonnes muettes celles qui sont historiques (*dompter*<lat. *domitare*) de celles qui sont étymologiques (*compter*<lat. *computare*). De même, en l'absence de transcription phonétique (que l'on serait d'ailleurs bien en peine d'établir pour cette époque), il reste compliqué d'identifier les lettres muettes de celles qui ne le sont pas, ou celles qui ont une valeur diacritique de celles qui n'en ont pas – une même variante pouvant d'ailleurs avoir plusieurs valeurs, le <s> de *teste* étant à la fois diacritique (longueur de la voyelle) et étymologique (<lat. *testa*).

Deuxièmement, l'analyse automatique du système graphique d'un texte ne résulte pas (encore) de l'opérationnalisation des typologies construites par les linguistes que nous venons de présenter, mais de l'analyse *ex post* des observations faites par la machine. En effet : comment automatiser l'analyse de la variation graphique ? Notre approche est de comparer une transcription aussi diplomatique que possible avec sa version normalisée : une fois les deux versions alignées au niveau des tokens (≈ les mots) et des caractères, il suffit d'observer des lieux de variations (*estoit* vs *était* : retrait de <s> et ajout de l'accent aigu + retrait de <o> et ajout de <a>) pour y détecter des grandes tendances, comme détaillé ailleurs (Gabay *et al.*, 2022). Cette approche a plusieurs implications importantes.

1. Tout ne peut être analysé. Les distinctions fines entre lettres étymologiques et les lettres historiques que nous avons préalablement mentionnées passent ainsi inaperçues, mais aussi tous les cas où le français contemporain a conservé une particularité que nous aimerions observer (par exemple le <g> de *doigt*<lat. *digitus*).
2. L'analyse par alignement produit un très grand nombre d'occurrences, inexploitables sans post-traitement. Pour remédier à ce problème, nous utilisons la partie d'analyse de l'outil ABA (Poinhos, 2020), qui permet de regrouper sous une même étiquette l'alignement d'une suite de caractères consécutifs, différente dans la version originale et dans la version normalisée du corpus parallèle. Ces regroupements forment des règles, faites principalement des alignements de caractères distincts les plus fréquemment observés, laissant de côté les cas les plus rares.
3. Ces règles doivent être déterminées à l'avance, en analysant un corpus dit *gold*, c'est-à-dire entièrement transcrit et normalisé à la main – il s'agit dans ce cas du corpus $\text{FREEM}_{\text{norm}}$ (Gabay *et al.*, 2022). De taille conséquente et construit avec soin, ce corpus n'est cependant pas représentatif du français du XVII^{e} s., et introduit donc un biais dans l'analyse.
4. C'est seulement dans un second temps que l'état de l'art relatif à l'étude de la variation graphique a été pris en compte pour nommer certaines de ces règles en fonction des phénomènes linguistiques qu'elles représentent, éventuellement en fusionnant certaines règles ayant une formulation similaire, par exemple lettres

ramistes pour les modifications entre les lettres <i> et <j> ou entre <u> et <v>, ou pour la règle <x>/<z> → <s>.

5. À l'inverse, le code de détection des règles d'analyse est capable d'assurer que chaque caractère du texte obtenu après alignement (intégrant éventuellement des caractères vides dans la version originale, en cas d'alignement avec un mot plus long en version normalisée, ou vice-versa, du type *apôtre* pour *apostre*) est concerné au maximum par une seule règle d'analyse. Cette propriété permet de considérer quantitativement des unions de règles avec la garantie qu'elles s'appliquent de manière disjointe, ce qui permet de cumuler les nombres d'occurrences de chaque règle pour obtenir le nombre d'occurrences de la combinaison des règles.

Le résultat prend la forme d'environ soixante-dix règles[7]. Pour les raisons précédemment évoquées, nous n'y retrouvons pas toutes celles relevées par J.-Ch. Pellat et Cl. Vachon, ni leur finesse d'analyse. Notre approche informatique offre cependant des nouvelles perspectives, en ouvrant la question des allographes (<ſ> → <s> ou <ß> → <ss>), ou en intégrant des cas qui appartiennent à la lexicologie avec, par exemple, celui des soudures (*quoi que* → *quoique*).

Nos résultats sont loin d'être parfaits, mais certains défauts peuvent être aisément corrigés à moyen terme – notre étude a ainsi valeur de preuve de concept. Dans les cas silencieux où le français contemporain a conservé des consonnes muettes déjà présentes au XVII^e^ s. (*dompter*, *doigt*, etc.), il est possible de déduire la présence de ces lettres d'une comparaison entre toutes les occurrences d'une même forme à disposition : s'il est impossible d'identifier le <p> en surcharge dans *dompter* car la forme contemporaine est identique, la présence de la graphie *domter* dans le même corpus permet de déduire par triangulation la présence d'une lettre muette. Pour les occurrences éliminées du fait de leur faible fréquence mais dont l'importance linguistique est connue, il est possible de les réintroduire dans le système. Concernant le contexte gauche et droit de chaque zone de variation, nos relevés ne sont pas encore assez précis, mais il reste néanmoins possible d'en tenir compte pour affiner les règles que nous utilisons et proposer des sous-règles rendant mieux compte des réalités de la langue classique. Enfin, le corpus *gold* peut être élargi et mieux balancé afin de corriger les fréquences utilisées pour établir les règles.

À côté de ces défauts, une telle méthode a plusieurs avantages, et notamment trois. Le premier, précédemment évoqué, est de laisser la machine construire, tout du moins pour partie, le savoir elle-même : le chercheur n'intervient que pour analyser des résultats détectés mécaniquement, offrant un second regard, différent et donc complémentaire. Ce faisant, et c'est notre second point, il est possible d'étudier des phénomènes inédits, parfois considérés comme peu pertinents mais dont la forte fréquence peut être exploitée statistiquement sur de grandes quantités de données (comme le <s> long, témoin du matériel typographique à disposition dans les ateliers, dont l'influence sur la graphie a été étudiée par Biedermann-Pasques, 1992 : 92). Ce type d'analyse de masse est grandement facilité par les nouveaux outils, comme la reconnaissance optique de caractères (manuscrits comme imprimés) et la normalisation automatique (Bawden *et al.*, 2022). Enfin, en dépit des limites dont nous avons fait état, nous pensons que notre approche permet de confirmer des analyses d'ordre philologique, concernant des micro-variations à l'intérieur des textes, comme dans les imprimés de Nicolas Boileau Despréaux.

4. Cas d'étude : Boileau et ses *Œuvres diverses* de 1674

Afin de valider notre approche computationnelle, nous proposons une analyse de plusieurs corpus liés aux *Œuvres diverses du sieur D****, publiées par Boileau chez Thierry en 1674. J.-Ch. Pellat a en effet noté dans les pièces publiées pour la première fois par l'édition Thierry (*Art poétique*, *Lutrin*, *Traité du Sublime*) quelques particularités typiques des habitudes manuscrites de Boileau, particularités dont les textes réimprimés dans le même ouvrage (*Satires*) sont dépourvus (Pellat, 2001 ; cf. tab. 3)[8].

Le cas est intéressant car il est complexe. Plus que les simples *scriptae*, voire leur mélange, ce que nous nous proposons de détecter sont donc des traces du système graphique de Boileau, marqué par des traits particulièrement modernes (comme la finale avec accent de la P5 <-és>), se maintenant discrètement en diasystème avec la langue du copiste, plus conservateur dans sa pratique (la finale de la même P5 <-ez>, sans accent). En d'autres termes : sommes-nous capables, sur la base des règles détectées automatiquement, de classifier correctement les différentes parties de l'ouvrage en fonction de leur statut d'inédit ou non, et donc de confirmer l'hypothèse de J.Ch. Pellat ? Question subsidiaire : la comparaison de cet imprimé avec la documentation manuscrite autographe confirme-t-elle cette classification ? Pour cette enquête, nous mobilisons trois corpus liés à l'imprimé de 1674 :

- un corpus BOILEAUNORM constitué d'extraits de l'ouvrage obtenus par reconnaissance optique de caractères et correction manuelle par le logiciel *eScriptorium* (Kiessling *et al*., 2019) ainsi que la version normalisée manuellement de chaque extrait ;
- le même corpus avec cette fois une normalisation automatique par un modèle de traduction automatique décrit ci-dessous, BOILEAUNORMAUTO.
- le corpus BOILEAUNORM+, constitué du corpus BOILEAUNORM augmenté de la transcription d'une version manuscrite autographe d'une pièce versifiée de Boileau, l'*Ode à la prise de Namur*, envoyée à Racine en 1693 (Paris, BNF, Fr. 12886, f°126), et sa normalisation (manuelle).

Tableau 3. Description du corpus du cas d'étude : pages et vues Gallica des sous-parties du recueil Œuvres diverses du sieur D***, et des extraits corrigés

Sous-Corpus	Œuvre	Pages	Vues Gallica	Vues transcrites
P1	Discours au roi	1-6	7-12	7-12
P2 et P3	Satires 1 à 9	7-67	13-73	13-18 et 19-22
	Discours sur la satire	69-75	75-81	-
P4 et P5	Epistres au roi 1 à 4	77-99	83-105	83-90 et 91-93
P6	L'Art poétique	103-142	109-148	109-117
P7	Le Lutrin, chants I à IV	149-178	163-192	163-171
P8	Traité du sublime et du merveilleux	NP-91	195-292	204-211

4.1 La méthode de normalisation automatique

Le modèle de normalisation automatique correspond au meilleur modèle proposé par Bawden *et al.* (2022). Il repose sur une approche de traduction automatique statistique

dite « phrase-based ». Le modèle a été entrainé avec MOSES (Koehn *et al.*, 2007), l'outil de référence pour ce type de modèles. L'objectif d'un tel modèle de traduction est de déterminer, étant donnée une séquence de mots donnée en entrée, la séquence de mots qui constitue la traduction la plus probable de la séquence d'entrée. Dans notre cas, la séquence d'entrée est une phrase en français classique et la traduction recherchée est son équivalent respectant les conventions de français contemporain, la norme choisie dans ce travail. Étant donnée une phrase d'entrée, la façon dont on mesure combien une traduction candidate est probable (c'est-à-dire combien il est probable qu'elle soit correcte) s'appuie sur plusieurs modules qui assignent chacun un score à cette traduction candidate. Les deux principaux scores sont fournis par :

- un *modèle de traduction*, c'est-à-dire une table de correspondance entre séquences de *n* mots dans la langue source et des traductions possibles de ces séquences, chacune de ces correspondances étant munie d'une probabilité estimée à partir d'un grand corpus parallèle d'entraînement ;
- un *modèle de langue* de type *n*-gramme qui modélise la probabilité qu'une traduction candidate donnée soit une phrase correcte dans la langue cible, ce modèle ayant été entraîné sur un grand corpus de phrases en langue cible, qui peut être le versant cible du corpus parallèle mentionné au point précédent.

Le modèle est entraîné sur le corpus $\text{FREEM}_{\text{norm}}$ (Gabay, 2022), composé de textes de genres variés du XVII^e^ s. et leurs normalisations manuelles. Une étape de post-traitement est utilisée, qui s'appuie un lexique de français contemporain, le Le*fff* (Sagot, 2010). Son rôle est de normaliser les mots qui, dans la sortie du modèle, sont inconnus du lexique mais sont très proches (aux diacritiques près et modulo quelques autres changements très simples) d'exactement un mot connu du lexique, afin d'éviter toute situation d'ambiguïté. Ainsi, si le modèle produit *etait*, il sera normalisé par cette étape de post-traitement en *était*, seul mot connu du Le*fff* et suffisamment proche d'*etait*.

Nous avons évalué ce modèle à l'aide d'une métrique d'exactitude lexicale. Cette métrique consiste, sur des textes pour lesquels une normalisation manuelle existe, à évaluer le pourcentage des mots de la version normalisée produite automatiquement qui est identique à leur équivalent dans la version normalisée manuellement. Ce modèle s'appuyant sur une approche statistique s'est avéré plus performant que des modèles alternatifs utilisant des méthodes plus récentes qui s'appuient sur les réseaux de neurones ainsi que des modèles par règles bien plus simples mais déjà performants. Sur le corpus décrit dans cette section, notre modèle atteint une exactitude lexicale de 97,4%.

4.2 Les résultats obtenus

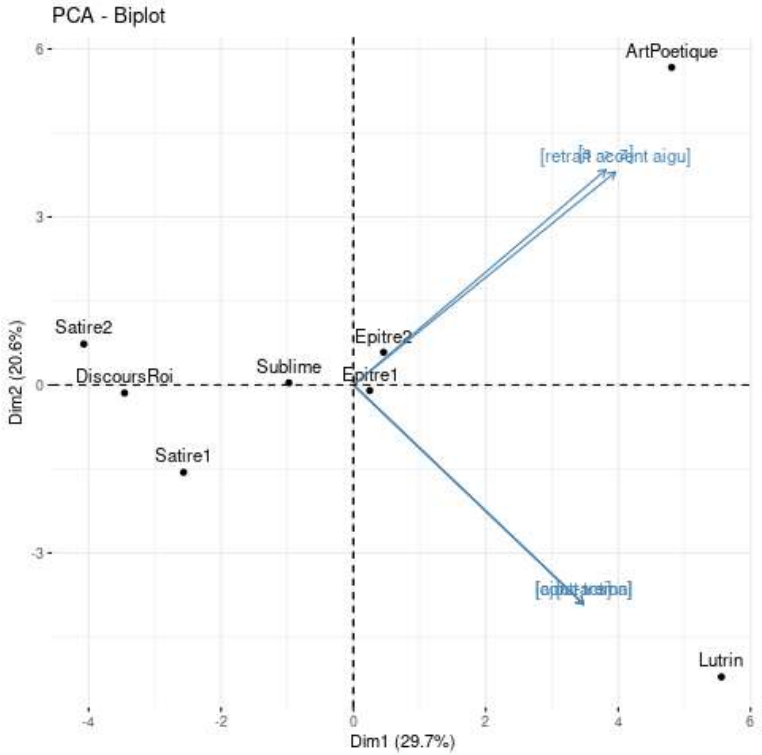
Pour chacun des deux corpus d'extraits de textes de Boileau, nous comptons le nombre moyen d'occurrences de chaque règle détectée pour 10 000 mots. Nous visualisons ces données dans un graphique à deux dimensions à l'aide d'une analyse en composantes principales (ACP). Le principe de ce traitement de données vise à réduire le nombre de dimensions (initialement, ce nombre de dimensions est égal au nombre de règles de normalisation détectées dans les corpus, soit une quarantaine) à deux, en trouvant un plan en deux dimensions qui maximise la variance de l'axe horizontal et de l'axe vertical, permettant ainsi de projeter sur ce plan les points correspondant à chaque texte. Cette réduction, qui implique nécessairement une déformation que l'on doit contrôler[9], permet de retenir une partie de la variance jugée suffisante pour indiquer

les grandes tendances. Le résultat final prend la forme d'un plan où chaque texte est caractérisé par l'application de toutes les règles de normalisation.

Si nos hypothèses de travail se trouvent confirmées, nous devrions obtenir deux types de résultats pour nos deux questions de recherche. D'une part, lors de la comparaison des textes inédits avec ceux republiés dans l'édition de 1674, les deux groupes devraient former des clusters distincts. D'autre part, lors de la comparaison de l'édition de 1674 avec la documentation manuscrite, cette dernière devrait se retrouver plus proche des textes inédits que des textes republiés.

Nous obtenons ainsi les illustrations 1 (a) et (b) pour les normalisations respectivement manuelle et automatique. On y constate, une séparation le long de l'axe horizontal entre, à gauche, deux satires et le *Discours au roi* qui les précède dans l'imprimé de 1674, et à leur droite, les extraits de textes publiés pour la première fois à cette occasion. L'organisation le long de l'axe horizontal de ces textes auparavant inédits est très similaire dans les deux illustrations, avec de gauche à droite :

- le Traité du sublime et du merveilleux ;
- deux Épitres (numérotées 1 et 2) ;
- l'Art poétique ;
- le Lutrin.



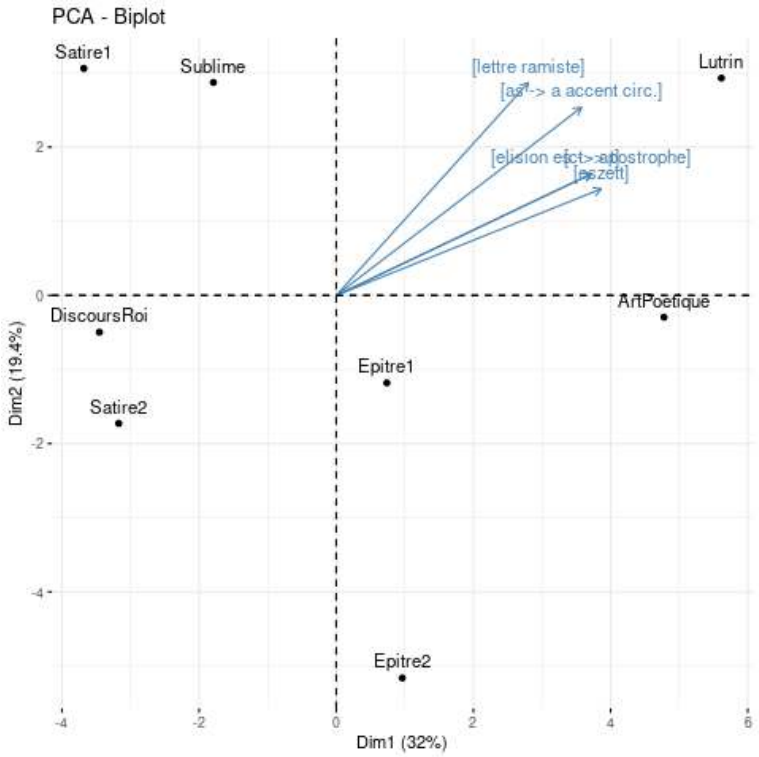


Illustration 1. Analyses en composantes principales des sous-corpus parallèles obtenus par normalisation manuelle (a) et par normalisation automatique (b).

Il est appréciable que les corpus parallèles issus de la normalisation manuelle et de la normalisation automatique fassent tous deux ressortir le même signal le long de cet axe principal – en d'autres termes, le passage par une automatisation de la tâche de normalisation pour l'analyse des graphies, en dépit de son imperfection[10], fonctionne. Ce signal met en évidence des règles correspondant aux phénomènes remarqués par J.-Ch. Pellat :

> - l'ajout d'accent aigu et la transformation du <s> en <z> (cf. tab. 4a) sont les deux règles d'ABA, dont les vecteurs sont orientés vers la droite dans cette ACP, qui correspondent le plus souvent au morphème de la P5 : *Aimé**s*** → *Aime**z*** dans *L'Art poétique*, *demeuré**s*** → *demeure**z*** dans *L'Épitre 1* ou encore *Tené**s*** → *Tene**z*** dans *L'Épitre 2*.
> - le <y> calligraphique, dont le vecteur est orienté vers la gauche dans cette ACP (cf. tab. 4a et 4b), correspond à la transformation finale de <y> en <i> comme dans *a**y*** → *a**i*** dans les deux satires et dans le *Discours au Roi*, ou encore dans les mots *pourquo**y***, *vra**y*** et *quo**y*** ; notons toutefois que l'on retrouve ces trois dernières formes aussi dans le *Traité du sublime et du merveilleux*, ce qui n'avait pas été noté par J.-Ch. Pellat et peut contribuer à expliquer la position horizontale de ce texte entre les textes précédemment publiés et ceux également publiés pour la première fois en 1674.

À l'inverse, la présence des formes finales en <ez> au lieu de <és> dans les textes déjà publiés, évoquée par J.-Ch. Pellat, n'est pas particulièrement détectée : la règle d'ABA correspondante, <x>/<z> → <s> correspond dans *affame**z***, *Beaute**z*** et *forme**z*** dans les *Satires*, ou à *oppose**z***, *accuse**z*** et *entraîne**z*** dans le *Traité du sublime et du merveilleux*.

Tableau 4. Sélection des principales règles obtenues par normalisation manuelle (a) et par normalisation automatique (b).

Règle	Dimension 1	Dimension 2	Forme originale	Forme normalisée

Eszett	0.87230002	0.32438134	preße	presse
ct → t	0.83993067	0.37013350	fruict	fruit
as → â	0.80742658	0.57248286	hastée	hâtée
lettre ramiste	0.63218216	0.64782818	aueugle	aveugle
Lettre calligraphique	-0.82788867	-0.03373438	luy	lui

Règle	Dimension 1	Dimension 2	Forme originale	Forme normalisée
retrait accent aigu	0.71176615	0.68381941	allés	allez
s → z	0.68579621	0.69104233	allés	allez
retrait du tréma	0.62523163	-0.70413309	obeïr	obéir
lettre ramiste	0.62523163	-0.70413309	Jront	Iront
c → s	0.62523163	-0.70413309	offence	offense
Lettre calligraphique	-0.76927685	0.10965883	luy	lui

En analysant le corpus BOILEAUNORM+[11], on constate que l'analyse en composantes principales fournie dans l'illustration 2 sépare :

- selon le premier axe (horizontal), les pièces imprimées à gauche et le manuscrit de l'*Ode à Namur* à droite ;
- selon le deuxième axe (vertical), on retrouve dans la partie supérieure le manuscrit et les pièces publiées pour la première fois dans l'imprimé de 1674 (*Epitres*, *Art poétique*, *Traité du sublime*, *Lutrin*), retenant respectivement tout ou une partie du système graphique de Boileau. Les pièces déjà publiées (première et seconde *Satires*, *Discours au Roi*) dans la partie inférieure.

On note que l'ensemble des deux axes explique plus de la moitié (57 .5%) de la variance. Parmi les règles de normalisation particulièrement associées au manuscrit de l'*Ode à Namur*, on trouve l'ajout de l'accent grave (ex. : 14 occurrences de <a> → <à>, de *déjà* → *déjà),* la désagglutination des proclitiques (ex. : *davancer*, *leffroi*, *senfuit*), le changement du morphème flexionnel verbal de la P1 <-y> en <-is> (ex. : *vo**y***), le remplacement du <-s> final par <-z> (ex. : *Voié**s***, *Approché**s***), souvent cumulé avec le retrait de l'accent aigu (ex. : *Voi**é**s*, *Approch**é**s*), le retrait du <c> muet dans le digramme <ct> (*fai**ct***, *sain**ct***), l'ajout de tréma (*héro**ï**que*), le remplacement de <i> par <y> (*Déplo**i**és*, ***I**pres*, *Vo**i**és*) et la réintroduction de géminées (ex. : *sou**ff**le*) ou leur réduction (ex. : *De**ff**enseur*).

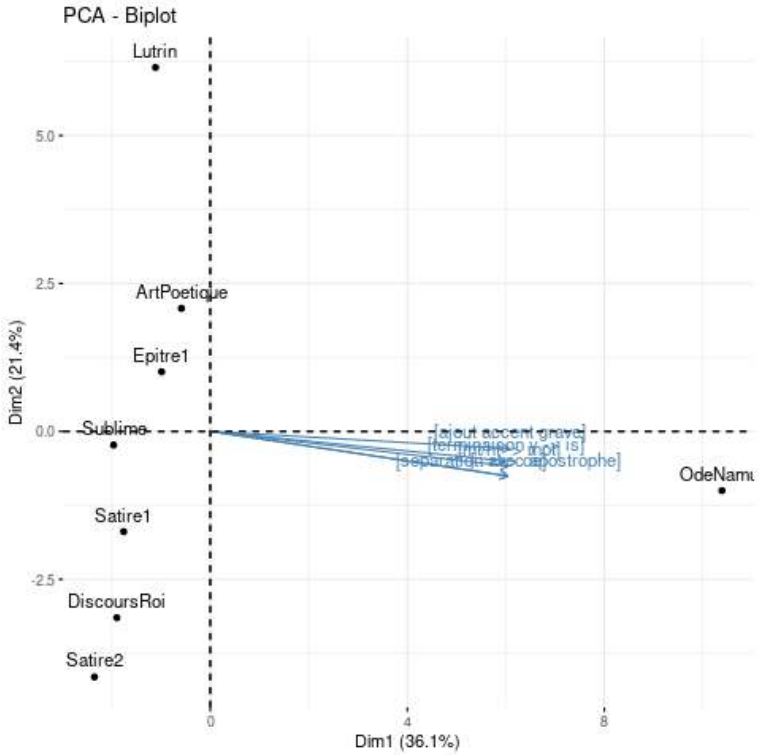


Illustration 2. Analyse en composantes principales des règles de normalisation détectées sur le sous-corpus parallèle BOILEAUNORM+.

Tableau 5. Valeur des règles de normalisation détectées sur le sous-corpus parallèle BOILEAUNORM+.

Règle	Dimension 1	Dimension 2	Forme originale	Forme normalisée
séparation avec apostrophe	0.97908083	-0.12233126	Dun	*D'un*
terminaison y → is	0.97069152	-0.08051458	ſçay	sais
mt/nt → mpt	0.98061917	-0.09536899	domta	dompta
ajout accent grave	0.98517666	-0.04718862	ou	où

5. Conclusion

Notre analyse computationnelle des écrits de Boileau tend à démontrer que les écrits publiés pour la première fois en 1674 partagent des particularités qui, pour partie, se retrouvent dans la documentation autographe, tandis que les textes publiés précédemment en partagent d'autres. Ces particularités détectées automatiquement au moyen de règles de normalisation recoupent les observations de J.-Ch. Pellat, mais par des moyens partiellement différents : le nombre d'observations mobilisées pour la constitution des clusters est en effet bien plus conséquent, et si le résultat souffre d'évidentes limites techniques, il a l'avantage d'une plus grande systématicité. Le travail de J.-Ch. Pellat nous permet donc de valider notre méthode, autant que nos analyses confirment les hypothèses de notre collègue.

Le système graphique de Boileau est marqué par des traits modernes que l'on retrouve assez nettement dans la documentation manuscrite, et partiellement dans les imprimés. Dans ce cas précis, qui n'a pas forcément de valeur générale, l'orthographe actuelle dont parle Cl. Vachon et que l'on observe dans l'imprimé de 1674 s'explique donc partiellement par le mélange de différents systèmes graphiques. Le mécanisme de copie produit logiquement la superposition de deux systèmes linguistiques (Segre, 1976), celui du prote et celui de l'auteur, comme le montre l'étude des *editiones principes* de Boileau (Pellat, 2001).

Il convient donc de revaloriser l'imprimé, *proxy* (imparfait) d'une documentation manuscrite souvent disparue au XVII^e^ s., mais qui peut conserver les vestiges d'états antérieurs autographes qu'il est possible de retrouver – à la manière des médiévistes qui, depuis fort longtemps, distinguent dans leur éditions la langue de l'auteur de celle du copiste, et ont pour ce faire développé des techniques dont les modernistes gagneraient certainement à s'inspirer. L'argument souvent entendu selon lequel la langue des imprimés n'est pas celle de l'auteur, et peut donc sans état d'âme être modernisée, est factuellement faux, au moins dans le cas que nous avons étudié ici, et probablement dans d'autres que notre méthode pourrait contribuer à débusquer. Pour reprendre l'expression de Stephan Elspaß, une *Sprachgeschichte von unten* doit donc en partie se faire « par en haut », certains textes « publics » témoignant en filigrane d'usages « privés » – si tant est que cette distinction soit faisable (Altman, 1992).

Cette porosité encore mal étudiée entre les manuscrits et les imprimés de la première modernité est une trace importante de la lente convergence de l'écrit, dont la première étape a été parfaitement décrite par Y. Greub (2007). Selon ce dernier, la phase médiévale de préstandardisation est (en partie) portée par un phénomène de neutralisation mécanique des marques diatopiques du fait du processus de copie. Comme Y. Greub le suggère à la fin de son article, nous pensons que le processus continue pendant la première modernité, ou plus précisément se rejoue d'une manière différente à cause de l'imprimé. Le copiste est désormais le prote d'imprimerie qui appose plus ou moins parfaitement son système graphique par-dessus celui de l'auteur, qui peut néanmoins survivre en diasystème. Comme l'a parfaitement démontré J.-Ch. Pellat, cette survivance ne dure pas dans le temps et disparaît dès la deuxième copie, c'est-à-dire, dans un monde d'imprimés, d'une nouvelle édition.

Le nombre de copies imprimées étant logiquement supérieur à l'unique version autographe de l'auteur dans un schéma de transmission textuelle standard[12], l'impact de l'imprimé est bien plus fort que celui du manuscrit : il accélère le processus de convergence. On retrouve ainsi de possibles traces de cette accélération dans la correspondance de Sévigné, avec une opposition nette entre la mère et la fille autour de l'usage de l'apostrophe et de l'emploi des lettres ramistes (Gabay, 2020). Cela expliquerait finalement l'importance supérieure d'usages individuels sur des *scriptae* moins nettement définies, qui se maintiendraient pendant la première modernité sur un mode dégradé, comme celles des Anciens et des Modernes, avant leur effacement final devant un standard linguistique régissant le français savant au cours du XVIII^e^ siècle.

BIBLIOGRAPHIE

ALTMAN, J. G., 1992, « Espace public, espace privé : La politique de la publication de lettres sous l'Ancien Régime », *Revue Belge de Philologie et d'Histoire*, 70, p. 607–23, <https://doi.org/10.3406/rbph.1992.3834>.

BADDELEY, S., 1996, « Tentatives de standardisation orthographique et typographique chez les imprimeurs français au XVIe siècle », dans M. Tavoni, P. Dini, J. Flood *et al.* (éds), *Italia ed Europa nella linguistica del Rinascimento: confronti e relazioni: atti del Convegno internazionale: Ferrara, Palazzo Paradiso, 20-24 marzo 1991*, Ferrare, Franco Cosimo Panini, t. 1, p. 287-300.

BAWDEN, R., POINHOS, J., KOGKITSIDOU, E., GAMBETTE, Ph., SAGOT, B., GABAY, S., 2022, « Automatic Normalisation of Early Modern French », dans *LREC 2022 Proceedings*, European Language Resources Association, <https://hal.inria.fr/hal-03540226>.

BERGERON-MAGUIRE, M., 2019, « Du Poitou en Louisiane : édition et notes à partir de la correspondance d'une peu lettrée (1802-1803) », *Géolinguistique*, 19, <https://doi.org/10.4000/geolinguistique.1530>.

BIEDERMANN-PASQUES, L., 1992, *Les Grands Courants orthographiques au XVII^e^ siècle et la formation de l'orthographe moderne*, Tübingen, M. Niemeyer, < https://doi.org/10.1515/9783110938593>.

BOSSUET, J. B., 1890, *Œuvres Oratoires de Bossuet*, J. Lebarq. (éd.), Lille/Paris, Desclée, De Brouwer et Cie, 7 vols.

BRAZEAU, St., LUSIGNAN, S., 2004, « Jalon pour une histoire de l'orthographe française au XIVe siècle. L'usage des consonnes quiescentes à la chancellerie royale », *Romania*, 122, p. 444-467, < https://www.persee.fr/doc/roma_0035-8029_2004_num_122_487_1333 >.

BUFFIER, Cl., 1709, *Grammaire Françoise Sur Un Plan Nouveau : Pour En Rendre Les Principes Plus Clairs et La Pratique Plus Aisée...* Paris, Nicolas Le Clerc, Michel Brunet, Leconte et Montalant, <http://gallica.bnf.fr/ark:/12148/bpt6k50481x>.

CATACH, N., 2001, *Histoire de l'orthographe française*, Paris, Champion.

Cahiers de remarques sur l'orthographe françoise pour estre examinez par chacun de Messieurs de l'Academie, avec des observations de Bossuet, Pellisson, etc. Publiés avec une introduction, des notes et un table alphabétique, 1863, Ch. J. Marty-Laveaux (éd.), Paris, Jules Gay, <https://books.google.ch/books?id=u5Y5AQAAIAAJ>.

DUVAL, Fr., 2015, « Les éditions de textes du XVII^e^ siècle », Dans *Manuel de La Philologie de L'édition*, dans D. Trotter (éd.), Berlin/Boston, De Gruyter, p. 369–94, <https://doi.org/10.1515/9783110302608-017>.

EDER, M., 2013, « Mind your corpus: systematic errors in authorship attribution », *Literary and Linguistic Computing*, 28, p. 603-614, <https://doi.org/10.1093/llc/fqt039>.

FUMAROLI, M., 2001, « Les Abeilles et les araignées », dans *La Querelle des Anciens et des Modernes, XVII^e^ - XVIII^e^ siècles*, Paris, Gallimard.

GABAY, S., 2014, « Pourquoi moderniser l'orthographe ? Principes d'ecdotique et littérature du XVII^e^ siècle », *Vox Romanica*, 73, p. 27–42, <https://elibrary.narr.digital/article/99.125005/vox201410027>.

GABAY, S., 2020, « La naissance de Marie-Blanche de Grignan. Notes sur la mise en page de la polyphonie sévignéenne », *Acta Litt&Arts*, 13 (*Les discours rapportés en contexte épistolaire (XVIe-XVIIIe siècles)*, <https://hal.archives-ouvertes.fr/hal-01900042>.

GABAY, S., 2022, *FreEM-corpora/FreEMnorm: FreEM norm Parallel corpus* (version 1.0.0), Zenodo, <https://doi.org/10.5281/zenodo.5865428>.

GABAY, S., BAWDEN, R., GAMBETTE, Ph., POINHOS, J., KOGKITSIDOU, E., SAGOT, B., 2022, « Le changement linguistique au XVII^e^ s. : nouvelles approches scriptométriques », dans F. Neveu, S. Prévost, A. Steuckardt, G. Bergounioux et B. Hamma (eds), *Actes Du 8^e^ Congrès Mondial de Linguistique Française*, https://doi.org/10.1051/shsconf/202213802006.

GOEBL, H., 1975, « Qu'est-ce que la scriptologie ? », *Medioevo Romanzo*, 2, p. 3-43.

GOSSEN, C., 1967, *Französische Skriptastudien, Untersuchungen zu den nordfranzösischen Urkundensprachen des Mittelalters*, Vienne, R. M. Rohre.

GREUB, Y., 2007, « Sur un mécanisme de la préstandardisation de la langue d'oïl », *Bulletin de La Société de Linguistique de Paris*, 102, p. 427–32, <https://doi.org/10.2143/BSL.102.1.2028211>.

JEJCIC, F., 2017, « Écritures dialectales (1865-1997), en marge de l'histoire de la langue », dans A. Kristol (éd.), *La Mise à l'écrit et ses conséquences : Actes du troisième colloque 'Repenser l'histoire du français', Université de Neuchâtel, 5-6 juin 2014*, Tübingen, A. Francke, p. 211-236.

KIESSLING, B., TISSOT, R., STOKES, P., STÖKL BEN EZRA, D., 2019, « EScriptorium: An Open Source Platform for Historical Document Analysis », dans *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Institute of Electrical and Electronics Engineers, t. 2, p. 19–19, <https://ieeexplore.ieee.org/document/8893029>

KOCH, St., 2013, « Sobre el contacto del leonés con el castellano en la Edad Media. Estudio preliminar de ocho documentos de San Pedro de Eslonza (1241-1280) », dans E. Herrero et C. Rigual (eds), *Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas*, Berlin/Boston, De Gruyter, t. 8, p. 595-607, < https://doi.org/10.1515/9783110300031.595>.

KOEHN, Ph., HOANG, H., BIRCH, A., CALLISON-BURCH, Ch., FEDERICO, M. BERTOLDI, N., COWAN, B., *et al.* 2007, « Moses : Open Source Toolkit for Statistical Machine Translation », dans *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 177–80, Prague, Association for Computational Linguistics, <https://aclanthology.org/P07-2045>.

PELLAT, J.-Ch., 1992, « Corneille et la modernisation de l'orthographe au XVII^e^ s. », *Le Français Moderne*, 60, p. 161–70.

PELLAT, J.-Ch., 1994, « Norme et variation orthographique au XVII^e^ s. », dans *Rencontres linguistiques en pays rhénan 5/6*, Strasbourg, Université des sciences humaines de Strasbourg, p. 245–60.

PELLAT, J.-Ch., 2001, « L'orthographe des poètes du XVII^e^ s. : Boileau et La Fontaine », dans J.-Ch. Pellat, Cl. Buridant et G. Kleiber (eds), *Par monts et par vaux : Itinéraires linguistiques et grammaticaux. Mélanges de linguistique Générale et Française offerts au Professeur Martin Riegel pour son soixantième anniversaire par ses collègues et amis*, Louvain/Paris, Peeters Publishers, p. 305–22.

POINHOS, J., 2020, *ABA (Alignment-Based Approach)*, version 1, <https://github.com/johnseazer/aba>.

REMACLE, L., 1948, *Le Problème de l'ancien wallon*, Liège, Presses universitaires de Liège. <http://books.openedition.org/pulg/338>.

RIFFAUD, A., 2007, *La Ponctuation du théâtre imprimé au XVIIe siècle*, Paris, Droz.

SAGOT, B., 2010, « The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French », dans *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, European Language Resources Association, <https://hal.inria.fr/inria-00521242>.

SEGRE, C., 1976, « Critique textuelle, théorie des ensembles et diasystème », *Bulletins de l'Académie Royale de Belgique*, 62, p. 279-292 < https://www.persee.fr/doc/barb_0001-4133_1976_num_62_1_55259>.

VACHON, Cl. H., 2010, *Le Changement linguistique au XVI^e^ s : Une étude basée sur des textes littéraires français*. Strasbourg, ELiPhi, Éditions de linguistique et de philologie.

NOTES

1. Notons que si *stricto sensu* l'orthographe renvoie à une manière correcte d'écrire une langue, elle peut aussi correspondre « aux diverses façons d'écrire en interaction avec la norme » (Jejcic, 2017 : 233). L'ambiguïté introduite par ce sens second nous paraît problématique au XVII^e^ s., et notamment sa seconde moitié qui voit l'Académie s'occuper abondamment des questions orthographiques.

2. Cf. aussi H. Goebl qui se propose d'étudier les *scriptae* « avant leur unification sous l'hégémonie socio-linguistique d'une orthographe prépondérante. [...] l'objet de nos investigations se présente encore sous l'aspect d'une certaine malléabilité » (Goebl 1975 : 3).

3. Nous prolongeons donc la remarque d'H. Goebl, selon qui « Toute scripta est donc un continuum hybride et composite, offrant tout un faisceau de traits typiques régionaux (ou diatopiques), archaïsants, innovateurs, et bien d'autres encore » (Goebl 1975 : 3).

4. On retrouvera une analyse plus précise dans Pellat (1994).

5. Cette inconsistance reste relativement limitée comparée à celle présente dans la langue médiévale, comme le montre la lecture des textes de l'époque.

6. À des fins de lisibilité, la description proposée par Cl. Vachon est considérablement simplifiée. De nombreuses sous-sections existent pour chaque catégorie.

7. Le détail des règles est disponible à l'adresse suivante : https://github.com/johnseazer/aba/blob/master/aba/utils/modern.py dans la fonction *find_diffs*.

8. Denis [II] Thierry est marchand-libraire et imprimeur : la composition est donc effectuée dans ses ateliers. Le processus de copie (texte copié, pratiques d'atelier...) nous reste en revanche inconnu : seule l'étude de la langue nous permet de proposer quelques hypothèses.

9. Le pourcentage d'information retenu est égal à la somme des deux pourcentages associés à l'abscisse et l'ordonnée de chaque plan. Le résultat obtenu pour chacune des ACP *infra* d'environ 50% est considéré comme tout à fait correct pour une réduction de 40 à 2 dimensions, et autorise l'interprétation des résultats. Un résultat bien inférieur (par ex. 20%) ne serait pas suffisant pour tirer des conclusions solides.

10. Chaque erreur de normalisation entraîne la détection ou l'absence de détection de certaines règles, ce qui affecte pour partie l'expérience. Pour partie seulement, car nous savons qu'un niveau raisonnable de bruit n'invalide pas le résultat final (Eder, 2013).

11. Pour cette expérience nous ignorons deux règles : <ſ> → <s> et <&> → <et>. La distinction <ſ>/<s> est en effet particulièrement difficile à faire dans les manuscrits, et l'esperluette est une pratique d'imprimeur. En conservant ces deux règles nous accentuerions inutilement le clivage manuscrit/imprimé alors que nous cherchons à évaluer l'influence du premier sur le second.

12. Si ce schéma standard est la norme, il existe évidemment des modèles alternatifs. L'auteur peut ainsi superviser l'établissement du texte imprimé, ou le témoin de base du copiste peut être un allographe, s'il est par exemple l'œuvre d'un secrétaire à qui l'on a dicté le texte.

RÉSUMÉS

L'abandon des systèmes graphiques au profit de l'orthographe contemporaine dans l'écrasante majorité des éditions de textes du XVII^e^ siècle. a fait disparaître leur richesse graphématique. Cette dernière est logiquement restée en grande partie méconnue, et donc sous-exploitée, en dépit de son rendement interprétatif certain. En privilégiant une approche pratique, sur corpus, plutôt que théorique, et en s'appuyant sur une recatégorisation des différents systèmes concurrents à cette époque en *scriptae* françaises, nous nous proposons de poser les bases d'une étude scriptométrique de la langue classique, s'intéressant à l'analyse de documents singuliers, manuscrits comme imprimés.

The use of contemporary spelling rather than old graphic systems in the vast majority of current editions of 17th century French texts has the unfortunate effect of masking their graphematic richness. Such valuable information has remained concealed and therefore under-exploited, despite the potential it holds in terms of analysis. By favouring a practical corpus-based approach, rather than a theoretical one, and by relying on a recategorisation of the various competing systems at that time in French scriptae, we propose the foundations of a scriptometric study of the classical language, focusing on the analysis of specific documents, both manuscripts and old prints.

INDEX

Mots-clés : graphématique, philologie computationnelle, scripta, orthographe, traitement automatique des langues, linguistique historique
Keywords : graphematics, computationnal philology, spelling, orthography, natural language processing, historical linguistics

AUTEURS

SIMON GABAY

Université de Genève

PHILIPPE GAMBETTE

Université Gustave Eiffel, UMR 8049 LIGM

RACHEL BAWDEN

INRIA

BENOÎT SAGOT

INRIA