Humanités numériques et computationnelles appliquées à l'étude de l'écrit ancien

Peter Anthony Stokes

EPHE
Sciences historiques et philologiques

Annuaire de l'École pratique des hautes études. Section des sciences historiques et philologiques

155 | 2024
Annuaire de l'EPHE, section des Sciences historiques et philologiques (2022-2023)

Humanités numériques et computationnelles appliquées à l'étude de l'écrit ancien

Peter A. Stokes

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/ashp/7703
DOI : 10.4000/11twe
ISSN : 1969-6310

Éditeur
Publications de l'École Pratique des Hautes Études

Édition imprimée
Date de publication : 1 septembre 2024
Pagination : 522-528
ISSN : 0292-0980

Référence électronique
Peter A. Stokes, « Humanités numériques et computationnelles appliquées à l'étude de l'écrit ancien », *Annuaire de l'École pratique des hautes études (EPHE), Section des sciences historiques et philologiques* [En ligne], 155 | 2024, mis en ligne le 13 juin 2024, consulté le 18 juin 2024. URL : http://journals.openedition.org/ashp/7703 ; DOI : https://doi.org/10.4000/11twe

HUMANITÉS NUMÉRIQUES ET COMPUTATIONNELLES APPLIQUÉES À L'ÉTUDE DE L'ÉCRIT ANCIEN

Directeur d'études : M. Peter A. Stokes

Programme de l'année 2022-2023 : *Vers une paléographie transversale : la fonction d'écriture et comment la décrire*.

La problématique principale de cette année est la continuation de notre travail sur la modélisation de l'écriture, en revenant sur la question de sa fonction. Bien qu'une telle discussion puisse sembler très théorique, elle est cruciale pour savoir comment partager les données et les méthodes numériques entre différents systèmes d'écriture. Comme nous l'avons évoqué dans les rapports précédents[1], il existe de nombreux exemples d'individus maitrisant des écritures très différentes, telles que l'arabe, le grec et le latin, le chinois et le sogdien, ou encore le chinois et l'hébreu. Si nous voulons pouvoir étudier et comprendre ces pratiques, nous avons besoin d'un langage et d'un modèle conceptuel communs. Ceci est d'autant plus important quand nous voulons publier, partager et lier des données provenant de différents projets et bases de données, comme c'est l'objectif de Biblissima+. Une telle approche presque typologique s'est avérée en pratique difficile à mettre en œuvre. Les tentatives précédentes sont l'œuvre de chercheurs issus principalement de la linguistique qui se concentrent donc (de manière compréhensible) implicitement sur l'écriture contemporaine, et en particulier sur la fonction dénotative directe de l'écriture en tant que vecteur d'informations linguistiques[2]. Il manque donc toujours une vision historique et diachronique sur la terminologie et les définitions des termes, ainsi qu'une discussion plus complète sur la fonction connotative de l'écriture. L'objectif de notre conférence cette année a donc été de (ré)élaborer certaines définitions et certains concepts clés, en tant qu'une première étape afin d'aborder des questions plus générales.

Graphie

Le terme le plus facile à définir est sans doute celui de graphie. Définie par Davis comme « a unique instance of [a letter] as it appears on the page »[3], ce mot fait référence à une *marque physique* sur une surface créée par une personne par le biais d'une *action*

1. P. A. Stokes, « Humanités numériques et computationnelles appliquées à l'étude de l'écrit ancien », *Annuaire. Résumés des conférences et travaux, 154e année, 2021-2022*, Paris, EPHE-PSL, SHP, 2023, p. 517-524, et P. A. Stokes, « Humanités numériques et computationnelles appliquées à l'étude de l'écrit ancien », *Annuaire. Résumés des conférences et travaux, 153e année, 2020-2021*, Paris, EPHE-PSL, SHP, 2022, p. 529-531.
2. La plupart des discussions sur ce sujet sont fortement influencée par les théories structuralistes, notamment les ontologies récentes telles que le Cidoc-CRM (voir n. 4 ci-dessous). Cette approche s'est avérée très utile et est suivie ici, bien que d'autres approches soient évidemment possibles et susceptibles de s'avérer productives.
3. Tom Davis, « The Practice of Handwriting Identification », *The Library: The Transactions of the Bibliographical Society*, 8 (2007), p. 251-276 à la p. 255 ; comparez également Dimitrios Meletis, « Types of Allography », *Open Linguistics*, 6 (2020), p. 249-266 à la p. 252, et Dimitrios Meletis, et

délibérée qui peut ensuite être *interprétée* par un lecteur comme un signe porteur d'information linguistique. À cet égard, il s'agit clairement d'un exemple (plus strictement, d'une sous-classe) de E25 « Human-Made Feature », telle que définie par le modèle de référence conceptuel du CIDOC[4].

Allographe

Contrairement à la graphie, l'allographe est beaucoup plus difficile à définir. On a proposé de le définir comme « an accepted version of [a] grapheme », ou « a conditioned or free variant of a character »[5], mais ces définitions sont très imprécises, surtout lorsqu'elles s'appliquent à des systèmes d'écriture non européens. Il est clair que l'allographe est « émique » dans le sens où il s'agit d'un idéal imaginé ou construit et non d'une marque concrète sur une surface. Il s'agit néanmoins d'une forme visuelle qui représente un graphème et qui est destinée à être interprétée comme tel. À cet égard, il semble évident que l'allographe appartient à la classe E36 « Visual Object » tel que défini dans le modèle de référence conceptuel du CIDOC, à savoir « the intellectual or conceptual aspects of recognisable marks, images and other visual works », ce qui réfère non pas à des instances spécifiques d'une image mais à son « underlying prototype » qui « remains uniquely identifiable ... independent[ly] of ... [its] visual support »[6]. En développant ces points, nous pouvons donc dire qu'un allographe est un prototype visuel qui est *partagé* et qui peut (et doit) être *appris*. Il existe dans un contexte culturellement déterminé de variation paradigmatique libre ou conditionnée. Il représente un graphème qui peut être identifié au moins en partie par la relation syntagmatique avec d'autres allographes. L'allographe lui-même n'est pas porteur d'informations linguistiques, mais il a une signification *connotative* à travers sa forme visuelle, car l'utilisation de différents allographes est elle-même significative. En effet, les allographes n'ont de sens que dans le contexte d'un système (d'écriture) qui établit les variantes conditionnées ou libres qui sont autorisées[7], et qui établit les allographes qui sont non marqués (c'est-à-dire les formes « défaut » sans connotation particulière) ou marqués (formes qui sont différentes et donc portent une signification connotative). Par exemple, dans le contexte de cet article, un allographe « défaut » (non marqué) est R, mais d'autres allographes incluent $\mathfrak{R}$ ou $\mathbb{R}$, et ceux-ci peuvent avoir des connotations différentes dans des contextes différents. Ceci est la base d'une blague dans *Astérix et les Goths*, par exemple, où tout ce qui est dit dans la langue

Christa Dürscheid, *Writing Systems and Their Use: An Overview of Graphonlinguistics*, Berlin, 2022, p. 63.

4. Chryssoula Bekiari, George Bruseker, Martin Doerr, Christian-Emil Ore, Stephen Stead, et Athanasios Velios (dir.), *Definition of the CIDOC Conceptual Reference Model*. Version 7.1.1. ICOM-CIDOC, 2021.
5. T. Davis, « The Practice of Handwriting Identification », p. 255, et Peter T. Daniels, William Bright (éd.), *The World's Writing Systems*, New York, 1996, p. XXXIX, respectivement.
6. Bekiari *et al.*, *Definition of the CIDOC Conceptual Reference Model*, p. 82.
7. Meletis note que la variation libre des allographes est rare et constitue une exception aux règles d'un système d'écriture (D. Meletis, « Types of Allography », p. 254), mais en fait cette variation est souvent observée dans l'écriture historique, comme plusieurs formes de **a** et de **s** trouvées dans le minuscule d'Angleterre du onzième siècle, pour lequel voir, par exemple, Peter A. Stokes, *English Vernacular Minuscule from Æthelred to Cnut, circa 990 – circa 1035*, Cambridge, 2014 (Publications of the Manchester Centre for Anglo-Saxon Studies 14).

des Goths et écrit dans un style Fraktur (marqué), et Astérix et les Gaules répond dans le style non marqué « par défaut »[8]. Cela pourrait suggérer une autre caractéristique des allographes, analogue à celle des graphèmes, à savoir qu'un allographe est la plus petite unité contrastive significative au niveau *connotatif* d'un système d'écriture.

Un aspect mal défini de l'allographe est de savoir si deux formes similaires (ou à quel point d'un spectre plus ou moins continu de formes) sont des allographes différents. Meletis répond en partie à cette question en définissant différents niveaux d'allographie[9], mais cette distinction s'applique mieux à l'imprimé et est beaucoup moins claire dans le cas de l'écriture manuscrite. En principe, la réponse dépend de la définition de l'allographe comme « accepté », ce qui signifie que la différence doit être reconnue comme distinctive par une communauté donnée. Un point qui n'est pas clair ici est de savoir si la définition se situe du point de vue du lecteur ou de celui qui écrit, c'est-à-dire, est-ce que la distinction des allographes repose sur la démonstration d'une conscience (plus ou moins) des formes allographiques de la part du scribe. Dans ce cas, l'allographe est effectivement une forme prototype qui joue un rôle dans la production de graphies. Par ailleurs, cette forme peut être acceptée par les lecteurs, qu'ils soient contemporains du scribe ou (par exemple) des paléographes experts modernes. Dans ce cas, l'allographe est construit sur la base des graphies observées dans un contexte donné. De même, on pourrait imaginer des allographes différents en fonction des groupes des personnes, comme une forme qui n'est reconnue et acceptée que dans un scriptorium donné, ou dans une orbite culturelle donnée (comme ceux qui écrivaient les minuscules anglo-carolines), ou à une plus grande échelle (les scribes de minuscules carolines plus généralement). Ce qui est clair, c'est que l'allographe ne peut être défini que dans le contexte d'un système donné : les allographes d'un système ne seront pas nécessairement des allographes dans un autre, par exemple parce qu'ils ne sont pas reconnus comme distinctifs, ou parce que des formes allographiques dans un système sont graphémiques dans un autre (comme *i* et *ı* qui sont graphémiques en turc mais pas en français). De même, une forme donnée peut correspondre à un graphème dans un système et à un graphème différent dans un autre système ; par exemple, la forme visuelle o peut représenter le latin /o/ (comme dans *hot*), le grec /o/ (comme dans *όσ*), le cyrillique /o/ (comme dans *область*), l'hébreu /o/ (comme dans *ום*), le chiffre 0, un signe de ponctuation en chinois et dans les écritures apparentées (〇), et ainsi de suite. Dans la pratique, le seul moyen de distinguer ces signes différents est le contexte, qui permet normalement d'identifier le système d'écriture et donc de limiter suffisamment les possibilités pour déterminer l'allographe. Cela soulève à son tour un autre aspect de l'allographe qui n'est pas clair dans la plupart des

8. Je remercie Paolo Monella pour cet exemple. Un autre exemple similaire est « Gehen Sie wählen! Andere tun es auch » (Yannis Haralambous et Martin Dürst. « Unicode from a Linguistic Point of View », dans Yannis Haralambous (éd.), *Proceedings of Graphemics in the 21st Century, Brest 2018*, Brest, 2019), p. 167-83 à la p. 138, entre autres).
9. D. Meletis, « Types of Allograpy ». En résumé, dans le système de Meletis, les allographes graphématiques correspondent à la définition typique de l'allographe, y compris celle présentée ici, à savoir des formes visuellement différentes qui représentent le même graphème, tandis que les allographes graphétiques opèrent à des niveaux intermédiaires entre les graphies et les allographes graphématiques pour regrouper des graphies qui se ressemblent ou qui ont la même structure et la même configuration de traits.

définitions : à savoir si une forme visuelle dans deux systèmes d'écriture différents correspond à un seul allographe ou à deux allographes différents : par exemple, le latin /o/ et le grec /o/ représentent-ils un ou deux allographes ? Un concept utile ici est la forme de base (« basic form » en anglais), qui se réfère à la forme visuelle indépendamment du système d'écriture, et la forme de base peut à son tour représenter un ou plusieurs allographes en fonction du système donné[10]. Le cas le plus complexe est celui de l'écriture manuscrite, où un ensemble donné de marques sur la page peut être interprété différemment, par exemple des séquences de minimes dans l'écriture gothique textura, ou même dans l'écriture moderne, des écritures très endommagées (par exemple les tablettes de Vindolanda), des écritures historiques qui ne sont pas encore totalement comprises, etc. Dans ces cas, l'auteur a (vraisemblablement) voulu que les marques correspondent sans ambiguïté à une série donnée d'allographes. La difficulté réside donc dans l'interprétation du lecteur. Si nous souhaitons modéliser cette interprétation du lecteur, nous devons alors tenir compte de la correspondance d'une graphie à plusieurs graphèmes possibles (ou même à plusieurs graphies, ou parties de graphies, dans le cas où la division des marques en graphies n'est pas claire), mais dans ce cas, l'hypothèse est qu'il existe une seule correspondance correcte de chaque graphie à exactement un allographe, même si cette correspondance peut ne jamais être identifiée dans la pratique.

Graphème

La définition du graphème est encore plus controversée que celle de l'allographe[11]. La discussion la plus approfondie à ce jour est probablement celle de Meletis. Ce dernier définit le graphème comme une unité *minimale* qui est *lexicalement distinctive* et qui *a une valeur linguistique* en ce qu'elle représente une ou plusieurs combinaisons de phonèmes, de morphèmes, etc.[12]. Au-delà, il est clair une fois de plus qu'un graphème doit être *appris* et *partagé*, et doit exister et fonctionner dans un *système* en relation syntagmatique avec d'autres graphèmes (c'est-à-dire dans le contexte d'une langue et d'un texte donnés). Un graphème peut être et est souvent représenté par plus d'une forme visuelle, à savoir différents allographes.

Cela nous amène à un autre point de discussion, à savoir la relation entre l'allographe et le graphème. Comme nous l'avons évoqué, un allographe est une représentation visuelle d'un graphème. On suppose normalement que chaque allographe représente un et un seul graphème, mais ce n'est pas nécessairement le cas. Par exemple, dans les cultures biscriptales telles que le grec et le latin, la même forme visuelle peut correspondre à deux graphèmes différents, un dans chaque système (par exemple le latin /o/ et le grec /o/)[13]. Là encore, le contexte linguistique permet normalement d'identifier

10. L'une des définitions les plus complètes de la forme de base est donnée par D. Meletis et C. Dürscheid, *Writing Systems and Their Use*, p. 63-66. Leur discussion est une version quelque peu mise à jour de D. Meletis, « Types of Allography », p. 252-53. Notez que pour Meletis, la forme de base est elle-même une forme d'allographe (graphétique), par opposition à ce qu'il appelle un allographe graphémique.
11. Pour une présentation de cette discussion, voir Dimitrios Meletis, « The Grapheme as a Universal Basic Unit of Writing », *Writing Systems Research*, 11 (2019), p. 26-49, aux p. 27-34.
12. D. Meletis, « The Grapheme », p. 26.
13. Peter A. Stokes, « Describing Handwriting in Context », *Digital Humanities 2022 Conference Abstracts*, Tokyo, 2022, p. 376-378.

l'allographe et donc le graphème, bien que des ambiguïtés puissent subsister, intentionnellement ou non, comme dans le cas des artistes modernes qui utilisent délibérément une forme visuelle donnée pour représenter simultanément deux graphèmes ou plus, ou des scribes qui utilisent les « mauvaises » formes, ce qui jette le doute sur les autres formes. Par exemple, si un écrivain utilise le grec /ϕ/ à la place du latin /f/ dans un mot tel que *ϕilius* (pour *filius*), il n'y a *a priori* aucun moyen de savoir si la deuxième lettre du même mot était destinée à être la lettre latine ou le *iota* grec.

Une autre question liée à la définition des graphèmes est celle des majuscules et des minuscules, pour les systèmes d'écriture qui font une telle distinction. Par exemple, on peut se demander si les majuscules et les minuscules (*a* et *A*, par exemple) constituent des graphèmes différents, ou s'ils sont des allographes différents d'un même graphème. Comme l'a souligné Coulmas, la réponse dépend de la langue, citant l'exemple de l'allemand *Wand* « mur » et *wand*, « blesser »[14]. Bien que moins évidente, une distinction très similaire existe également en anglais moderne, comme le montre la paire « It's brown! » et « It's Brown! », où *brown* est un adjectif dans le premier cas et un nom propre dans le second (vraisemblablement une M^me^ ou un M. Brown), la différence sémantique étant indiquée par l'utilisation ou non de la majuscule *B*. Cela s'applique également à d'autres formes graphiques telles que le soulignement ou l'italique, qui peuvent également indiquer des distinctions sémantiques et sont donc, conformément à la définition ci-dessus, graphémiques. Un exemple est celui des titres d'œuvres qui sont également des noms propres, comme Jane Eyre (versus *Jane Eyre*), ou l'exemple connu des spécialistes du vieil anglais qui travaillent sur la « dating *Beowulf* » mais vraisemblablement pas sur « dating Beowulf ». Dans ces cas, la forme italique se réfère bien sûr à l'œuvre littéraire, et l'alternative au personnage, et là encore la distinction sémantique est portée par l'italique, ce qui suggère à son tour que l'italique atteint le seuil nécessaire pour être considéré comme graphémique, au moins dans ce contexte (entre autres). Ceci est l'une des raisons pour lesquelles certains chercheurs utilisent le concept de graphème suprasegmental[15].

L'alignement du modèle conceptuel

L'étape suivante consiste à aligner nos termes aux modèles conceptuels existants, tels que le CIDOC-CRM et son extension CRMtex, ainsi que ceux d'autres projets tels que Archetype et IDIOMCAT[16]. Un tel alignement permet de mieux définir nos termes dans le contexte d'autres glossaires plus établis, et il permet également d'établir des correspondances entre les bases de données existantes, ce qui est l'un des objectifs de cette discussion, comme indiqué ci-dessus. Un extrait d'un tel alignement est fourni dans la fig. 1, et la fig. 2 montre un exemple de ce modèle appliqué au mot *post*, écrit avec les

14. Florian Coulmas, *The Blackwell Encyclopedia of Writing Systems*, Oxford, 1996, p. 173.
15. Voir, par exemple, Jacques Anis, « Pour une graphématique autonome », *Langue française*, 59 (1983), p. 31-44, à la p. 41.
16. Pour l'alignement des modèles, ainsi que CIDOC-CRM et CRMtex, voir Stokes, « Humanités numériques » (2023), p. 20-521. Pour IDIOMCAT, voir F. Diehr *et al.*, « Modellierung von Entzifferungshypothesen in einem digitalen Zeichenkatalog für die Maya-Schrift », dans A. Kuczera, T. Wübbena et T. Kollatz (éd.), *Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten*, Wolfenbüttel, Forschungsverbund Marbach Weimar, 2019, http://dx.doi.org/10.17175/sb004_002.

deux dernières lettres en ligature (ſt). Cette approche permet également de prendre en compte d'autres cas tels que *ꝑilius* pour *filius*, où le ꝑ peut être considéré comme un élément visuel représentant le graphème « lettre F latine ».

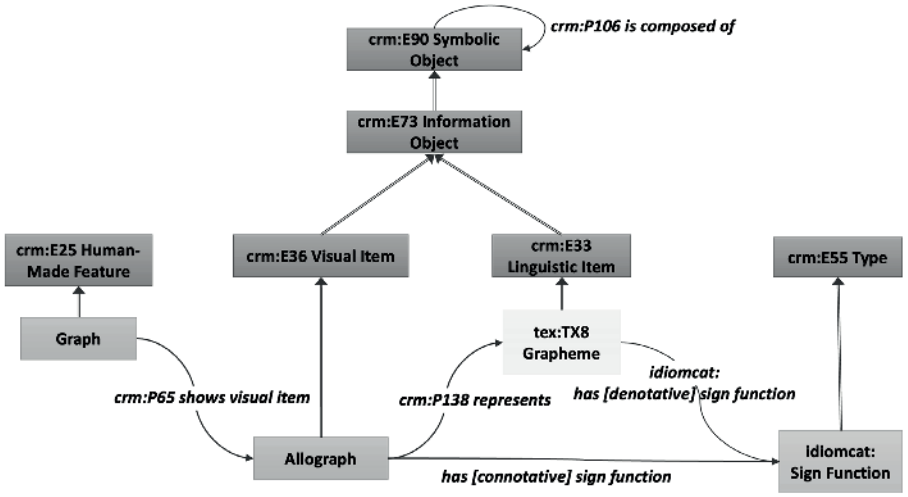

Fig. 1. — Extrait du modèle aligné avec ceux de Cidoc-CRM, CRMtex et Idiomcat.

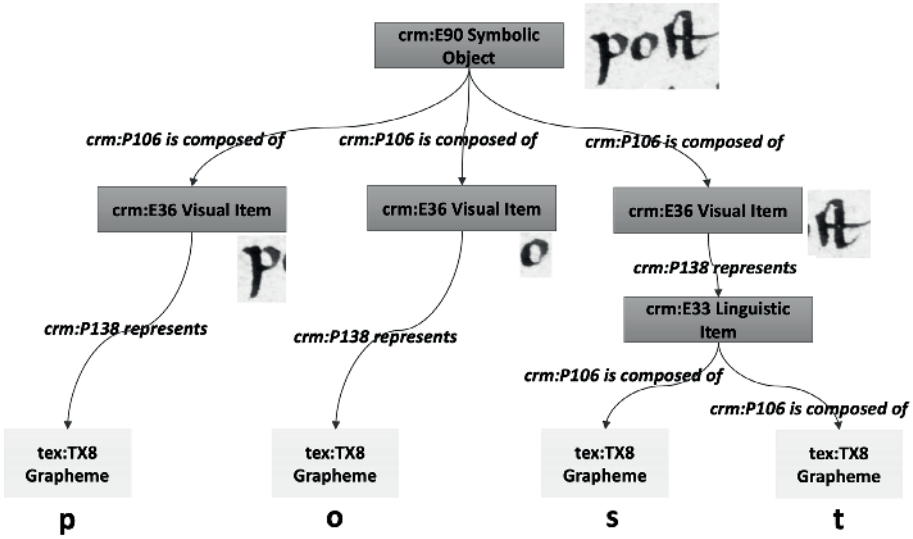

Fig. 2. — Modèle appliqué à l'exemple du mot *post* écrit en minuscule caroline.

Ce qui n'est pas encore discuté ici, c'est la fonction elle-même, et à cet égard, l'analyse de Klinkenberg et Polis est un exemple très utile[17]. Selon leur modèle, les fonctions

17. J.-M. Klinkenberg et S. Polis, *Les fonctions de l'écriture : un modèle général*, Liège, Collège Belgique, 2019, https://orbi.uliege.be/handle/2268/241566, ainsi que Stokes, « Humanités numériques » (2022). Des typologies plus détaillées sont fournies par des ontologies linguistiques spécialisées telles

des signes sont à la fois autonomes et relationnelles. Les fonctions autonomes comprennent normalement la transmission du son, du contenu ou des deux, et les fonctions relationnelles peuvent être lexicales, morphologiques, syntaxiques ou prosodiques. Ce modèle fonctionne au niveau dénotatif, tandis que la définition des fonctions connotatives est beaucoup plus difficile et doit être abordée ultérieurement.

que GOLD ou LEXINFO, bien qu'elles aient naturellement un objectif différent de celui de l'étude présentée ici. Voir Scott Farrar *et al.*, *General Ontology for Linguistic Description (GOLD)*, Linguist List (2010), https://linguistlist.org/gold/, entre autres.