Variability as a Key Factor For Understanding Medieval Scripts: the ORIFLAMMS project (ANR-12-CORP-0010)

Dominique Stutzmann

Variability as a Key Factor For Understanding Medieval Scripts: the ORIFLAMMS project (ANR-12-CORP-0010)*

Dominique Stutzmann

A revolution is taking place in the twenty-first century: as a consequence of globalisation and the increasing use of electronic devices for writing and reading, the human brain faces unprecedented challenges. New demands have emerged simultaneously: multilingualism; typing on a keyboard or writing letter-by-letter on touchscreen devices instead of with a pen; reading generic fonts on a screen instead of deciphering the handwriting of one's correspondents. In contrast, writing *by hand* involves the complete body (fingers, hand, arm, chest and head position, eye movements, breathing) in a three-dimensional space and multiplies physical, visual and tactile stimuli with many micro-controls, complex muscular and neural interactions. This develops rich neural influxes which are of immense significance for learning and memorising letter-forms and for writing ability. Neuroscientists have noted that when writing manually, muscle and brain memory speed up the process of memorisation and understanding – thereby proving the old proverb *Qui scribit, bis legit* (he who writes, reads twice) – and even allow one to read and understand a relative's handwriting faster than a typescript.[1] As a variable trace, and as part of the process of interpersonal communication, writing by hand plays a part in the development of one's ability to abstract, perceive and relate individual letter-forms to a set cognitive model. Thus, variability is a key factor in understanding script and its effects on the human mind.

Variability has always been a key concept in the humanities, as a factor of change and historical evolution as well as the core phenomenon between normativity, social control and individuality. Variability in written cultures is an issue for communication and literacy studies, linguistics, philology, history and palaeography, but also psychology and neuroscience. Philologists and linguists have faced variability in the contexts of phonetic evolution, linguistic regularisation and the emergence of orthography (or at least graphical systems).[2] Historians and palaeographers have addressed the issue in many ways, both in terms of 'Darwinian' evolution and progressive creation of new forms on the one hand, and stylisation, canonisation and 'Linean' taxonomy for more formal scripts on the other hand.[3] In connection with diplomatics, the concept of variability relates to those of validity, authenticity and attribution, since the identification of forgeries, spurious charters, autographs and scribal hands in books depends largely on how much variation is allowed within a chancery or scriptorium. The very notion of 'expertise' in palaeography relies on

* First submission on Sept. 21st, 2012; peer reviews and comments communicated on June 3rd, 2013; submission of revised version on Nov. 22nd, 2013; copy editing communicated on Oct. 20th, 2017; revised version submitted on Nov. 3rd, 2017. This author wish to thank the reviewers and editors for their valuable comments and input. As a consequence of the elapsed time since the second submission, some references and footnotes may be slightly outdated or incomplete: only the references directly linked to the ORIFLAMMS project have been updated. The conclusion summarizes the main results. For further information on the achievements of the ORIFLAMMS project, please refer to D. Stutzmann, "Compte-rendu final du projet ORIFLAMMS / ORIFLAMMS Final report", *Écriture médiévale & numérique* (1 April 2017) <http://oriflamms.hypotheses.org/1592>.

[1] J.-L. Velay, M. Longcamp, and M.-T. Zerbato-Poudou, *De la plume au clavier : est-il encore utile d'enseigner l'écriture manuscrite* (Marseille: Ministère de la Recherche, 2004); M. Longcamp, 'Etude comportementale et neurofonctionnelle des interactions perceptivo-motrices dans la perception visuelle de lettres. Notre manière d'écrire influence-t-elle notre manière de lire ?' (unpublished doctoral thesis, Université de la Méditerranée - Aix-Marseille II, Aix-en-Provence, 2003).

[2] A. McIntosh, 'Scribal Profiles from Middle English Texts', *Neuphilologische Mitteilungen*, 76 (1975), 218-35.

the idea that minimal variations are sufficient to distinguish one scribe from another.[4] The long term variations and similarities are, in turn, relevant to psychology and neuroscience as part of studies of clinical semiotics and research into signs, symbols and communication and their perception in different times and cultures.[5]

In order to better understand the challenges presented by modern technology, the variability of medieval scripts is an excellent field of research. Indeed, the social, mental, cultural and physical conditions which influence the evolution and specialisation of scripts may be analysed from societies with even more complex forms and a higher level of variability (multiple morphologies for one letter, abbreviations, etc.), as well as a high expectancy for multilingualism.[6] Moreover, medieval studies may bring new viewpoints to other disciplines, such as computer vision and optical character recognition by developing a 'graphical model',[7] and neuroscience through the concepts of the 'graphemic buffer' and 'allographic system'.[8]

Nevertheless the very notion of variability has not been clearly defined for medieval scripts and there is no acknowledged method to measure intra-scribal or inter-scribal variation. Since variability produces overlaps between categories, no commonly accepted criterion has been defined to distinguish several scripts one from one another, and the issue of taxonomy remains highly debated. We must always remember the variability and complexity of scripts when attempting to create an objective measure through computer vision. But what are the mechanisms to overcome these uncertainties and discrepancies? What kind of common reference points and criteria can we create to serve epigraphy, palaeography, diplomatics, linguistics, neuroscience and computer vision?

This article details how palaeography connects to the context and rationale to study the variability of scripts, and evidences how core features such as letter-forms can be analysed in relation to semiotic coherence, phonetics, linguistics and neuroscience. It suggests that variability would be precisely a question for "digital palaeography" to be addressed with the help of computer vision. In the last part, it describes the strategy that has been implemented within the research project ORIFLAMMS (Ontology Research, Image Feature, Letterform Analysis on Multilingual Medieval Scripts, 2013-2016) to created the necessary data sets and tools, especially to build a common reference corpus and a formal ontology which should serve as a touchstone to measure variability. The conclusion lists the achievements of the cross-domain research project, ranging from the reference corpus to new open source developments for text-image alignment and joint palaeographical and linguistic analysis, passing through new publications on digital humanities and interdisciplinary research.

---

[3] M. Stansbury, 'The Computer and the Classification of Script', in *Kodikologie und Paläographie im digitalen Zeitalter - Codicology and Palaeography in the Digital Age*, ed. by M. Rehbein, P. Sahle, and T. Schaßan (Norderstedt: BoD, 2009), pp. 237-49.

[4] L. Gilissen, L'expertise des écritures médiévales: Recherche d'une méthode avec application à un manuscrit du XI[e] siècle : le lectionnaire de Lobbes (Codex Bruxellensis 18018) (Gand: E. Story-Scientia, 1973).

[5] C. Sirat, J. Irigoin, and E. Poulle, *L'Écriture. Le Cerveau, l'œil et la main: Actes du colloque international du Centre national de la recherche scientifique* (Turnhout: Brepols, 1990); M. Cohen and J. Peignot, *Histoire et art de l'écriture*, (Paris: R. Laffont, 2005).

[6] A. Y. Durgunoğlu and L. T. Verhoeven, *Literacy Development in a Multilingual Context: Cross-Cultural Perspectives* (New York: Routledge, 1998); L. Peer and G. Reid, *Multilingualism, Literacy and Dyslexia: A Challenge for Educators* (London: David Fulton Publishers Ltd, 2000).

[7] A. Ciula, 'Digital Palaeography: Using the Digital Representation of Medieval Script to Support Palaeographic Analysis', *Digital Medievalist*, 1 (2005) <http://www.digitalmedievalist.org/journal/1.1/ciula/>; A. Ciula, 'The Palaeographical Method Under the Light of a Digital Approach', in *Kodikologie und Paläographie im digitalen Zeitalter - Codicology and Palaeography in the Digital Age*, ed. by M. Rehbein, P. Sahle, and T. Schaßan (Norderstedt: BoD, 2009), pp. 219-35

[8] B. Lechevalier, F. Eustache, and F. Viader, *Traité de neuropsychologie clinique* (Bruxelles: De Boeck Supérieur, 2008).

1. LEGIBILITY, LITERACY, COMPLEXITY, VARIABILITY

1.1. A COMMON OBJECT, SEVERAL DISCIPLINES IN HUMANITIES

Medieval scripts offer a number of productive perspectives from which to address the social, technical and cognitive issues of formal variability and multilingualism. Indeed, already established academic disciplines and 'auxiliary' sciences (epigraphy, palaeography, diplomatics, palaeotypography) may now engage in dialogue with linguistics, digital humanities, neuroscience and computer vision. The move has been made to deconstruct the barriers with neighbouring disciplines (especially history, palaeography, art history and communication studies) and each field may now widen its research interest and face new challenges.

Digital palaeography, for instance, has been applied with good results to historically narrow or graphically homogeneous corpora: one scribe or one manuscript (Christine of Pisan or Clara Hätzlerin; Hugo von Montfort's poems, ms. Heidelberg, cpg 329), one chancery or one specific time period (counts of Hainaut and Holland; English vernacular minuscule in the eleventh century or English vernacular manuscripts of the late Middle Ages).[9] Because of the limited time span being considered in these cases, the whole question of variability and dynamics and the internal synchronic coherence of graphic systems cannot easily be studied. In epigraphy as well, although epigraphy was a leader in the digital humanities (for example, with its Epidoc standard),[10] the study of letter-form or 'epigraphic palaeography' has been limited to dating inscriptions so that, on the one hand, there is no unified vocabulary to describe and study the letters and letter-forms, their systems and their relationships, and on the other hand, the plasticity of alphabetical signs in the three-dimensional epigraphic documentary source remains to be studied.

Variability does not concern only (digital) palaeography or epigraphy, but also linguistics. Indeed, linguistic phenomena correlate with the variability of scripts and are already being researched through corpus linguistics, especially for Old and Middle French. This research stream has been gaining experience since the late 1980s. Two main endeavours ought to be mentioned here: firstly, the Charrette project, initiated in Princeton by K. D. Uitti in 1989 and enhanced in collaboration with the French CESCM, was the first to publish online very finely tagged transcriptions of the complete textual tradition of a medieval romance, which allowed unique philological and linguistic research.[11] And secondly, the ICAR research team, gathering diachronic linguists and experts of text

---

[9] M. Aussems, 'Christine de Pizan and the Scribal Fingerprint: A Quantitative Approach to Manuscript Studies' (unpublished master's thesis, Universiteit Utrecht, Utrecht, 2007) <http://igitur-archive.library.uu.nl/student-theses/2006-0908-200407/UUindex.html>; M. Aussems, 'Christine de Pizan : the Scribal Fingerprint' (unpublished PhD thesis, University of Edinburgh, Edinburgh, 2011) < http://hdl.handle.net/1842/7789>; W. Hofmeister, A. Hofmeister-Winter, and G. Thallinger, 'Forschung am Rande des paläographischen Zweifels: Die EDV-basierte Erfassung individueller Schriftzüge im Projekt DAmalS', in *Kodikologie und Paläographie im digitalen Zeitalter - Codicology and Palaeography in the Digital Age*, ed. by M. Rehbein, P. Sahle, and T. Schaßan (Norderstedt: BoD, 2009), pp. 261-92; A. A. Brink, J. Smit, M. L. Bulacu, and L. R. B. Schomaker, 'Writer Identification Using Directional Ink-Trace Width Measurements', *Pattern Recognition*, 45 (2012), 162-71; J. Smit, 'Paleography and the Digital Middle Ages: Experiences with the Groningen Intelligent Writer Identification System (GIWIS)', (Third International MARGOT Conference, New York, 17 June 2010); J. Smit, 'Meten is weten? De toepassing van het Groningen Intelligent Writer Identification System (GIWIS) op Hollandse kanselarij-oorkonden, 1299-1345', *Bulletin de la Commission royale d'Histoire*, 176 (2010), 343-60; D. Scragg, A. Rumble, K. Powell, and D. Smith, *MANCASS C11 Database* (Manchester: Manchester Centre for Anglo-Saxon Studies, 2010), now offline, partially accessible thanks to Internet Archive <https://web.archive.org/web/20120425000331/http://www.arts.manchester.ac.uk/mancass/C11database/>; P. A. Stokes and others, *DigiPal: Digital Resource and Database for Palaeography, Manuscript Studies and Diplomatic* (London: King's College London, 2011) <http://www.digipal.eu> [accessed 1 Nov. 2013]; L. Mooney, S. Horobin, and E. Stubbs, *Late Medieval English Scribes* (York: Centre for Medieval Studies at the University of York, 2012).

[10] L. Anderson and others, EpiDoc: Guidelines for Structured Markup of Epigraphic Texts in TEI (Lexington: Stoa Consortium, 2007) < http://www.stoa.org/epidoc/gl/latest/> (accessed February 2012).

[11] C. Pignatelli, 'L'archive du Projet Charrette : huit témoins prêts à se livrer', in *Ancien et moyen français sur le Web : enjeux méthodologiques et analyse du discours (Actes du Colloque international Ottawa, 4-5 octobre 2002)*, ed. by P. Kunstmann, F. Martineau, and D. Forget (Ottawa: E. David, 2003), pp. 203-20; C. Pignatelli, 'De l'approche

metrics, developed the *Base de Français Médiéval* (http://bfm.ens-lyon.fr) which contains a wide range of typologically different texts and a rich semantic enhancement, and published the interactive digital edition of *Queste del saint Graal* on the TXM-WEB platform (http://portal.textometrie.org/bfm/?command=documentation&path=/GRAAL).[12] Yet, there is nothing comparable if we want to study variability and multilingualism in Latin texts. While it is widely accepted that the vernacular exhibits greater variability than Latin – the language of educated monks and clerks –[13] no comprehensive research has been undertaken nor published that confirms this. Recent studies, especially within the OMNIA project,[14] tend to present a more balanced picture of the linguistic reality. For a single lemma in Latin, there might be as many different forms as in vernacular languages, partly because of declension, partly because of the confrontation between different semantic and phonologic systems (especially for Latin words of vernacular origin), even if these are often hidden through editorial policies of regularisation. A particular case study should be the bilingual manuscripts, in which the variance of each language may be studied and compared in regard to linguistics, palaeography and image analysis. The trend of research on medieval bilingualism concentrates on countries for which multilingualism is an ethnic and social phenomenon or on translations,[15] and tends to neglect how

---

quantitative à l'interprétation philologique : en naviguant dans le Projet Charrette', in *XXV^e Congrès international de linguistique et philologie romane (Innsbruck, 3-8 septembre 2007)*, ed. by M. Iliescu, H. Siller-Runggaldier, and P. Danler (Berlin,: De Gruyter, 2010), VI, 289-96; *Chrétien de Troyes : Le chevalier de la Charette (Lancelot) : le 'Projet Charrette' et le renouvellement de la critique philologique des textes*, ed. by C. Pignatelli and M. C. Robinson (Tübingen: G. Narr Verlag, 2002); C. Pignatelli, 'Présence et fréquence de la ponctuation dans les manuscrits en vers du XIII^e siècle : les huit manuscrits du Chevalier de la Charrette au banc d'essai', in *Systèmes graphiques de manuscrits médiévaux et incunables français : ponctuation, segmentation, graphies, Actes de la Journée d'étude de Lyon, ENS-LSH, 6 juin 2005*, ed. by A. Lavrentiev (Chambéry: Université de Savoie, 2007), pp. 85-105; C. Pignatelli, 'Philological Perspectives on the Textual Corpus of The Charrette Project: A Rereading of the Transcriptions', in *Dame Philology's Charrette: Approaching Medieval Textuality through Chrétien's Lancelot, Essays in Memory of Karl D. Uitti*, ed. by G. Greco and E. Thorington (Tempe, AZ: Arizona Center for Medieval and Renaissance Studies, 2011), pp. 159-76.

[12] C. Guillot and others, 'Constitution et exploitation des corpus d'ancien et de moyen français', *Corpus*, 7 (2008) <http://corpus.revues.org/1495>; A. Lavrentiev, 'Pour une méthodologie d'étude de la ponctuation médiévale basée sur une approche typologique', in *Etudes sur le changement linguistique en français, communications du Colloque « Diachro 2 » , Paris, 15, 16, 17 janvier 2004*, ed. by Bernard Combettes and Christiane Machello-Nizia (Nancy: Presses Universitaires de Nancy, 2007), pp. 191-204; A. Lavrentiev, 'Typologie textuelle pour l'étude linguistique de manuscrits français médiévaux', in *Systèmes graphiques de manuscrits médiévaux et incunables français : ponctuation, segmentation, graphies, Actes de la Journée d'étude de Lyon, ENS-LSH, 6 juin 2005*, ed. by A. Lavrentiev (Chambéry: Université de Savoie, 2007), pp. 49-66; C. Marchello-Nizia and A. Lavrentiev, *Queste del saint Graal: Édition numérique interactive du manuscrit de Lyon (Bibliothèque municipale, P.A. 77)* (Lyon: École Nationale Supérieure, 2011-2017) <http://portal.textometrie.org/bfm/?command=documentation&path=/GRAAL>.

[13] P. A. Stokes, 'Rule and Variation in Eleventh-Century English Minuscule', in *Ruling the Script: Formal Aspects of Medieval Written Communication (Books, Charters and Inscriptions)*, ed. by S. Barret, D. Stutzmann, and G. Vogeler (Turnhout: Brepols, 2016), pp. 489-508.

[14] B. Bon, 'OMNIA – Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins', *Bulletin du centre d'études médiévales d'Auxerre | BUCEMA*, 13 (2009), pp. 291-92.

[15] R. K. Emmerson, 'Visualising the Vernacular: Middle English in Early Fourteenth-Century Bilingual and Trilingual Manuscript Illustrations', in *Tributes to Lucy Freeman Sandler: Studies in Illuminated Manuscripts*, ed. by K. A. Smith and C. H. Krinsky (London: Harvey Miller, 2007), pp. 187-204; M. P. Brown, 'Building Babel: The Architecture of the Early Written Vernaculars', in *Omnia disce: Medieval Studies in Memory of Leonard Boyle, O.P.*, ed. by A. J. Duggan, J. Greatrex, and B. Bolton (Aldershot: Ashgate, 2005), pp. 109-28; I. Larsson, *Pragmatic Literacy and the Medieval Use of the Vernacular: The Swedish Example* (Turnhout: Brepols, 2009); F. Duval and F. Vielliard, *Le miroir des classiques* (Paris: Ecole nationale des Chartes, 2007) <http://elec.enc.sorbonne.fr/miroir_des_classiques/>; G. Veysseyre, "Metre en roman' les prophéties de Merlin, voies et détours de l'interprétation dans trois traductions de l' 'Historia Regum Britannie", in *Moult obscures paroles : études sur la prophétie médiévale*, ed. by Richard Trachsler, Julien Abed and David Expert (Paris: Presses de l'Universite de Paris-Sorbonne, 2007), pp. 107-66; C. Galderisi and C. Pignatelli (dir.), *La traduction vers le moyen français : actes du IIe colloque de l'AIEMF, Poitiers, 27-29 avril 2006* (Turnhout: Brepols, 2007); Gervais de Tilbury, *Les traductions françaises des 'Otia Imperialia' de Gervais de Tilbury, par Jean d'Antioche et Jean de Vignay: édition de la troisième partie* (Genève: Droz, 2006); J.-P. Rothschild, *Les traductions au Moyen Âge et à la Renaissance: Cycle thématique de l'IRHT 2000-2001* (Paris-Orléans: IRHT - Aedilis, 2002), moved to

multilingualism and dialects may well interfere with palaeographic reality. There are no studies of the factors of variation or variability of scripts covering the Middle Ages and the different languages in use and, as a consequence, there is neither consensus nor the methodological tools to approach the notion of variability across the relevant disciplines.

1.2. GRAPHICAL SYSTEM: LETTER-FORM AMBIGUITIES AND LINGUISTICS

1.2.1 LETTER-FORM AMBIGUITIES

In parallel with the lack of historical overview, the inner structures of the graphical systems in the Middle Ages have not yet been traced, although they possess a momentum for computer-aided transcription tools. Human readers (palaeographers) develop strategies to overcome difficulties and formal ambiguities, which remain mostly unconscious.[16] The confusion of 'long **s**' for **f** is a well-known pitfall of OCR-systems; it has been noticed and partly overcome with language models, but there are many more issues for which human reading strategies, based on the graphical system, can be exported to computer software. For instance: in some handwriting from the thirteenth century, particular forms prove to be very ambiguous for a non-expert, such as **a**, **d**, and the ligature **ct** (Fig. 1) or the letter **s** written as a flat, modern **8** (Fig. 2). Similarly, the letters **S** and **g** are very alike (Fig. 1).

<https://irht.hypotheses.org/category/cycle-thematique/les-traductions-au-moyen-age-et-a-la-renaissance-2000-2001> [accessed 1 November 2017]; G. Hasenohr, 'Traductions et littérature en langue vulgaire', in *Mise en page et mise en texte du livre manuscrit*, ed. by H.-J. Martin and J. Vezin (Paris: Cercle de la librairie-Promodis, 1990), pp. 229-352; C. Galderisi, *Translations médiévales: Cinq siècles de traductions en français au Moyen Âge (XI^e^-XV^e^ s.): Étude et Répertoire*, 2 vols (Turnhout: Brepols, 2011); F. Fery-Hue, 'Tradlat. Traductions latines d'œuvres vernaculaires', 2008 <http://www.tradlat.org/> [accessed 6 October 2011].

[16] For formal difficulties in particular see S. Tarte, 'Digitizing the Act of Papyrological Interpretation: Negotiating Spurious Exactitude and Genuine Uncertainty', *Literary and Linguistic Computing*, 26 (2011), pp. 349-58.

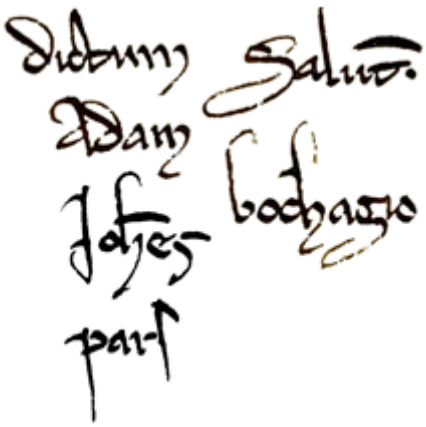

Fig. 1. Letters **a**, **ct** and **d**, **s** and **g** in 'dictum', 'salutem', 'Adam', 'Johannes', 'bochagio' and 'pars' (drawing after Arch. dép. Yonne, H 2402, dated 1225)

Fig. 2. Letters **a**, **d** and **s** in 'la', 'duc' and 'suis' (drawing after Arch. dép. Côte d'Or, 7 H 578, dated 1329)

Perception of the ambiguity is crucial, but its contextualisation is equally important. In the first example (Fig. 1), **a** or **s** are only ambiguous at the beginning of a word, since the script is characterised by its 'round **a**', 'long **s**' or 'sigma-shaped **s**' in other positions. This contextualisation is also needed for other forms. Indeed, some forms are ambiguous if we consider the complete history of the Latin alphabet, but their connections only appear in some particular scripts and, even within this reduced frame, a closer look at the letters will eventually show what makes the substantial difference within an example of the same handwriting; for instance, the position above the line (**S** and **g**) or the repartition of bold and thin strokes reflecting the actual *ductus* (**a**, **d** and **s**).

In order to gain a better understanding of the history of script and to overcome the issue of ambiguities, each graphical phenomenon ought to be analysed according to its context and its particular rules of appearance. Indeed, the script is simultaneously an idea and a form, which is written by a trained individual following a particular semiotic system; thus every one of these aspects (training, communication, form, idea, individual, social and system) may be an interpretative key. The core reason for a particular phenomenon may be internal to the script (interrelation between letters, as in some variants of Oeser's textualis), or refer to the materiality of the written document (end of line, end of page), or relate to the linguistic component (sound, word or sentence) or to the meaning of the text (names of persons and places, name of God).[17] Linguists intensively explore the graphic systems with regard to abbreviations, capitalization, punctuation and to the morphosyntactic phenomena (perception of word units or segmentation, syntax) or relationship between graphemes and phonemes,[18] but the hypothesis still needs to be confirmed because of the lack of an extensive and representative corpus.

1.2.2 Letter-forms and Phonemes

Exploring the morphologies, allographs and their correlation in the graphic chain should allow for a far more precise apprehension of the graphic system, and a better understanding of the interferences between graphic processes and linguistic phenomena. In particular, how the resemblance between distinct letters and the graphic variations may influence a word or morpheme recognition (by the human eye and brain) and reflect phonologic realities or semantic connections. Indeed, graphical variation may indicate different meanings for the same sound, or different sounds for

---

[17] D. Stutzmann, 'Paléographie latine et vernaculaire (livres et documents)', *Annuaire de l'École pratique des hautes études (EPHE), Section des sciences historiques et philologiques*, 144 (2013), pp. 115-28.

[18] N. Andrieux-Reix and S. Monsonégo, 'Écrire les phrases au Moyen Âge. Matériaux et premières réflexions pour une étude des segments graphiques observés dans des manuscrits français médiévaux', *Romania*, 115 (1997), pp. 289-336; N. Mazziotta, 'Traiter les abréviations du français médiéval. Théorie de l'écriture et pratiques d'encodage', *Corpus*, 7 (2008) <http://corpus.revues.org/1517>.

the same letter string. The latter is, for instance, present in the Beneventan script, in which the *ti*-ligature depends on the rendered sound.[19] This question still has to be extended to the other medieval scripts and letters. The example of letters *c* and *t*, whose shapes are very close, and which denote either distinct sounds (e.g. *ca*, *ta*), or the same phoneme (e.g. *ci*, *ti* in some words, but not all) is only one aspect, and perhaps, the easiest to explore. Marc Smith has suggested that the complete history of scripts has been influenced by the phonetic system of the associated written languages, as he notes that rarer phonemes are embodied in more peculiar forms: indeed, their respective frequencies have consequences for the graphical evolutions and, for instance, letters with ascenders and descenders correspond to the least used phonemes in pre-Carolingian Latin literature so that using a specific, outstanding form could serve as a reading-aid, a 'milestone' in visual perception.[20] Some historical examples also tend to prove that allographs may correlate phonetic phenomena and evolutions. For instance, in Périgueux, the town part called in Latin 'gressus' in the Middle Ages due to the presence of stairs, became progressively 'les greffes'.[21] The palaeographer cannot help thinking about the frequent confusion between 'long **s**' and **f** noted earlier.

1.2.3 Letter-form and Readability

The linguistic role of allographs and how they relate to word perception and semantics, remains very obscure. Modern French uses additional etymological letters to ensure the correct reading and understanding of words (e.g. phonetic [vɛːʀ] written '*vair*' [vair] / '*ver*' [worm] / '*verre*', '*verres*' [glass(es)] / '*vers*' [towards, verse, worms] / '*vert*', '*verts*' [green]; phonetic [so] written '*saut*', '*sauts*' [jump(s)] / '*sceau*', '*sceaux*' [seal(s)] / '*seau*', '*seaux*' [bucket(s)] / '*sot*', '*sots*' [fool(s)]). One might suppose that the positional allographs ensures the correct reading and understanding in a text: for example, the use of 'round **s**' at word end distinguishes close forms such as '*nef*' and '*nes*' or '*buef*' and '*bues*' in scripts in which 'long **s**' is least probable at the end of a word (cf. Fig. 3). Homophones and homographs are rarer in Latin than in Old French. Yet, diacritics were used whenever the writer felt a possible ambiguity ('*á mari*' against '*amari*')[22] and one might wonder if allographs and abbreviation forms were used for this purpose too. This could be explored on the basis of words like '*forum*' for which an abbreviation (half capital **r** with diagonal stroke) could have been used, but may have been avoided because it was rather used for a plural genitive -*orum*. The development of such reading aids and the balance between phonographic and logographic principles in writing and reading still requires an historic study.

Fig. 3. Positional allographs as diacritics.

In linguistics, the use of abbreviations is of particular interest. Some depict a one-to-one, bijective relationship between abbreviated form and expanded version (as for *p*-tilde for *pre*); others are to be interpreted according to the letters around them or with the whole word, esp. the *nomina sacra*. From a formal point of view, there is no impenetrable frontier, since the same signs are used for both sorts of abbreviations (tilde, semi-colon etc.). One

---

[19] E. A. Lowe, *The Beneventan Script: A History of the South Italian Minuscule* (Leiden: E.J. Brill, 1980).

[20] M. H. Smith, 'Les formes de l'alphabet latin, entre écriture et lecture', unpublished lecture, Collège de France (Paris), 14 October 2011 <http://www.college-de-france.fr/site/colloque-2011/symposium-2011-10-14-10h45.htm> [accessed 4 November 2011].

[21] M. Massénat, 'Périgueux médiéval et Renaissance' 2011 <http://www.cherveix-cubas.fr/cities/696/documents/d4h1846skh7y1.pdf> [accessed 19 August 2013].

[22] P. Bourgain, 'L'accent dans les manuscrits', in *Du copiste au collectionneur : mélanges d'histoire des textes et des bibliothèques en l'honneur d'André Vernet*, ed. by D. Nebbiai Dalla Guarda and J.-F. Genest, (Turnhout: Brepols, 1998), pp. 249-65.

abbreviation sign may have a different behaviour depending on language, time period, or region: the spatial variation, in particular, can give new insights in linguistic developments.[23]

Punctuation will not be excluded, for it is a valuable source on syntax perception and textual semantics and for the oral performance for which written texts are the only witnesses.[24] The use of punctuation is furthermore an important component of the scribal profile, although barely taken into account,[25] since there is no large scale digitised corpus presenting the original punctuation instead of the normalised one. In a future reference corpus, the larger part should be encoded at an allographic level including the punctuation so that one will be able to observe the evolution of the whole system. At a lower level, the blank space might also be explored as a punctuation mark: the graphic separation of words and morphemes is indeed a fundamental evolution of the Latin script and embodies the concrete perception of the word and marks the ability to read and reconstruct a meaning or reflects a grammatical sense. In Old French, the creation of graphically conjoined locutions (e.g. *par* + *mi*, *en* + verbs, *tres* + adjectives) is known, but not studied as such since there is no adequate corpus. A distinction that was probably perceived by medieval eyes is the differentiation of blank space between none/small/large, which does not convey any meaning to the modern eyes and has not been recognised yet, but ought to be observed and scientifically studied for the first time in this project.

The graphic chain and its formal variation, which does not exist in the modern typeset scripts, reflect the complex phenomenon of perception, and probably conceal precious information about reading strategies in the Middle Ages (e.g. whole language or phonics-based methods). As well as the graphic chain and the specialisation of allographs according to their place in the word, the separation of words matters to linguists, neuroscientists and palaeographers.

1.3. COMPUTER VISION AND VARIABILITY OF SCRIPTS AND LANGUAGES

1.3.1 DIGITAL PALAEOGRAPHY

In this context and with the development of computer vision and digital humanities, a series of goals seems to be obvious: (a) developing an automated optical character/handwriting recognition system; (b) creating a large collection of texts from the large digital libraries of medieval manuscripts, which would be the 'ground truth' and the training set for new recognition systems; (c) measuring and analysing linguistic variability; (d) measuring and analysing graphical systems by combining different techniques of pattern analysis and feature characterisation to cluster and categorise scripts. All four issues have to be addressed at the same time in a cross-fertilising approach and in an interdisciplinary and creative virtuous cycle. Analysing the graphical system opens new perspectives on consistency and variability of scripts. In this regard, the development of digital palaeography, the research in computer sciences and the development of new tools, such as computer-aided transcriptions, are, by no means, only a way to increase the speed of the inquiry or reduce the costs of expertise: these technologies build a new column in palaeographical analysis.[26] It is the unique method by which linguistic and literacy studies may take advantage of the new issues raised by the palaeographers and to take into account the materiality of texts, since handwritten text cannot be reduced to a unique transcription because the text is at least a two-dimensional object (matrix of pixels for digital photographs) and not only a linear text which can be fully represented by a string of letters.

---

[23] G. Hasenohr, 'Écrire en latin, écrire en roman : réflexions sur la pratique des abréviations dans les manuscrits français des XIIe et XIIIe siècles', in *Langages et peuples de l'Europe:. Cristallisation des identités romanes et germaniques (VII^e – XI^e siècle)*, ed. by M. Banniard (Toulouse: Presses universitaires du Mirail, 2002), pp. 79-110.

[24] C. Marchello-Nizia, 'Ponctuation et 'unités de lecture' dans les manuscrits médiévaux ou : je ponctue, tu lis, il théorise', *Langue française*, 40 (1978), 32-44; E. Llamas Pombo, 'Écriture et oralité : ponctuation, interprétation et lecture des manuscrits français de textes en vers (XIII^e-XV^e s.)', in *La linguistique française : grammaire, histoire et épistémologie, E. Aloinsi et al.,* ed by. E. Alonso, M. Bruña, and M. Muñoz (Sevilla: Grupo Analuz de pragmática, 1996), pp. 133-44.

[25] Pignatelli, 'Présence et fréquence' ; S. Barret, 'Reading the Charters is not Enough: Palaeography and the Diplomatist', Leeds, 11 July 2011.

[26] Hofmeister and others, 'Forschung am Rande'.

1.3.2 THE NEED FOR A LARGE REFERENCE DATASET

Three main factors explain the lack of a text recognition system for Latin medieval scripts and why there is no measurement of medieval scripts' variability: (a) the lack of a large training dataset covering the complete Middle Ages; (b) cursive scripts and ligatures; (c) and the immense variety and variability of scripts. Indeed, as for the lack of a large training dataset and ground truth, there is no reference corpus for the Central and Late Middle Ages which compares to reference corpora such as IAM MNISTRIMES, except two samples, one from the ninth century (Saint-Gall) and the other from the thirteenth century (Parzival).[27] As for the issue of cursivity and ligatures, the number of connex elements and the structural significance of connexity make it almost impossible to analyse the words at a letter-level or, at least, prevent attempts to do it at this stage, for there are only a few studies on the morphological consequences of cursivity and connexity on medieval scripts.[28] Word spotting techniques can only support script and linguistic analysis if the word matching is correct. This, in turn, is impossible due to the lack of a large ground-truth corpus on the one hand, and due to the great variety and variability of medieval scripts, on the other hand. The debates on taxonomy of medieval scripts and the overlapping of classes are of particular interest here.[29] One could easily imagine that one recognition system has to be trained for each script, but there is no consensus on what the different scripts are and how one already trained system should be propagated. As a consequence, systems can only be trained for one specific writer, rather than for a specific script and there is no measurement for either intra-scribal or inter-scribal variation.

As it is a prerequisite for all further research on medieval script, a reference dataset should be established. And since it is well known that there is no functional OCR for medieval manuscripts nor for incunabulas, even if some

---

[27] U. Marti and H. Bunke, 'The IAM-database: An English Sentence Database for Off-line Handwriting Recognition', *International journal on document analysis and recognition*, 5 (2002), 39-46; E. Indermühle, 'IAM Handwriting Database — Computer Vision and Artificial Intelligence' <http://www.iam.unibe.ch/fki/databases/iam-handwriting-database> [accessed 21 July 2012]; Y. LeCun and C. Cortes, 'The MNIST Handwritten Digit Database' <http://yann.lecun.com/exdb/mnist/> [accessed 21 July 2012]; 'RIMES database (Reconnaissance et Indexation de données Manuscrites et de fac similÉS / Recognition and Indexing of handwritten documents and faxes)', *RIMES*, 2012 <http://www.rimes-database.fr> [accessed 7 November 2013]; FKI: Research Group on Computer Vision and Artificial Intelligence IAM (University of Bern), 'Saint Gall Database — Computer Vision and Artificial Intelligence', *IAM Institut für Informatik und angewandte Mathematik*, 2012 <http://www.iam.unibe.ch/fki/databases/iam-historical-document-database/saint-gall-database> [accessed 18 October 2013]; FKI: Research Group on Computer Vision and Artificial Intelligence IAM (University of Bern), 'Parzival Database — Computer Vision and Artificial Intelligence', *IAM Institut für Informatik und angewandte Mathematik*, 2012 <http://www.iam.unibe.ch/fki/databases/iam-historical-document-database/parzival-database> [accessed 18 October 2013].

[28] E. Casamassima, *Tradizione corsiva e tradizione libraria nella scrittura latina del Medioevo* (Roma: Gela, 1988); S. Zamponi, 'Elisione e sovrapposizione nella littera textualis', *Scrittura e civiltà*, 12 (1988), pp. 135-76; I. Ceccherini, 'Tradition cursive et style dans l'écriture des notaires florentins (v. 1250-v. 1350)', *Bibliothèque de l'École des Chartes*, 165 (2007), pp. 167-85.

[29] Stansbury, 'The Computer'; G. I. Lieftinck, 'La nomenclature des écritures livresques du IX^{e} au XIII^{e} siècle', in *Nomenclatures des écritures livresques du IX^{e} au XVI^{e} siècle : premier colloque international de paléographie latine, Paris, 28-30 avril 1953*, ed. by B. Bischoff, G. I. Lieftinck, and G. Battelli (Paris: Édition du C.N.R.S., 1954), pp. 15-34; J. P. Gumbert, 'Nomenklatur als Gradnetz : ein Versuch an spätmittelalterlichen Schriftformen', *Codices manuscripti*, 1 (1975), pp. 122-5; J. P. Gumbert, 'A Proposal for a Cartesian Nomenclature', in *Essays Presented to G.I. Lieftinck*, IV, *Miniatures, Scripts, Collections* (Amsterdam: Van Gendt, 1976), pp. 45-52; A. Derolez, *The Palaeography of Gothic Manuscript Books from the Twelfth to the Early Sixteenth Century*. (Cambridge: Cambridge University Press, 2003); M. H. Smith, '[recension] Derolez (Albert). The Palaeography of Gothic Manuscript Books. From the Twelfth to the Early Sixteenth Century. Cambridge: Cambridge University Press, 2003', *Scriptorium*, 58 (2004), pp. 274-79; D. Stutzmann, 'Nomenklatur der gotischen Buchschriften: Nennen? Systematisieren? Wie und wozu? (Rezension über: Albert Derolez: The Palaeography of Gothic Manuscript Books. From the Twelfth to the Early Sixteenth Century. Cambridge u.a.: Cambridge University Press 2003.)', *IASLonline*, 2005 <http://www.iaslonline.lmu.de/index.php?vorgang_id=995>; G. Nicolaj, 'Questions terminologiques et questions de méthodes : autour de Giorgio Cencetti, Emanuele Casamassima et Albert Derolez', *Bibliothèque de l'École des Chartes*, 165 (2007), pp. 9-28; M. Palma, 'La definizione della scrittura nei cataloghi di manoscritti medievali', Viterbo, 5/03-2009; D. Stutzmann, 'Clustering of medieval scripts through computer image analysis: Towards an evaluation protocol', *Digital Medievalist* 10 (2016) <http://doi.org/10.16995/dm.61>.

developments have been made on enhanced letter-form modelling for set scripts,[30] and innovative projects associating pattern redundancy analysis and tools for 'human-assisted' transcription are now being developed,[31] one should use the various transcriptions and scholarly editions which are already available or make new transcriptions in which the text is directly aligned with the image, as in the T-PEN and TILE projects.[32]

1.3.3 THE NEED FOR A FORMAL ONTOLOGY

To envision and study variability at all the several abovementioned levels, we need an ontology of forms (an ontology formally represents knowledge as a set of concepts within a domain, and the relationships between pairs of concepts). The form ontology ought to provide the necessary basis to deepen our understanding of graphic systems and how much a script can be seen as a coherent system of signs and rules.[33] This ontology has already been dreamt of,[34] and recalls some developments of the SPI-System for Palaeographic Inspection,[35] but has never been achieved, although such a structural approach is of great importance for palaeography and has a particular relevance to linguistics and neuroscience.

Building an ontology of all available forms means not only creating a complete dictionary of glyphs and forms, but also creating an arborescence or, better, a graph of forms, and a vocabulary to describe them. This graph should avoid top down constraints and emerge from the raw data, building up classes, according to hypotheses on formal similarity which will have to be validated against the computational methods.

Neuroscience, linguistics and palaeography could also use this ontology. M. Smith underlined how likely it is that palaeographers are not sufficiently aware of systemic ambiguities or similarities.[36] These are obvious examples: if **C** and **G** were very similar in the Roman Cursive and in Uncial scripts, they became very distinctive as early as Half-uncial; on the contrary, letters **P** and **Q** became mirror-letters in Carolingian times. Less obvious is the systemic evolution and the relationship between ascenders, loops and descenders, for example. It is largely assumed that

---

[30] Ciula, 'Palaeographical Method'; A. Ciula, 'Modelli digitali di scrittura carolina', *Gazette du livre médiéval*, 45 (2004), pp. 27-38; G. Tomasi and R. Tomasi, 'Approche informatique du document manuscrit', in *Kodikologie und Paläographie im digitalen Zeitalter - Codicology and Palaeography in the Digital Age*, ed. by M. Rehbein, P. Sahle, and T. Schaßan (Norderstedt: BoD, 2009), pp. 197-218.

[31] Impact. Improving Access to Text, *OCR in Mass Digitisation Challenges between Full Text, Imaging and Language 6-7 April 2009, KB The Hague* (Den Haag: Koninklijke Bibliotheek, IMPACT Project Office, 2009); C. Kämmerer, 'Vom Image zum Volltext – Möglichkeiten und Grenzen des Einsatzes von OCR beim alten Buch', *Bibliotheksdienst*, 43 (2009), pp. 626-59; F. Rayar and J.-Y. Ramel, 'PaRADIIT Project', *PaRADIIT Project* <https://sites.google.com/site/paradiitproject/> [accessed 18 October 2013]; M. J. Castro Bleda, 'Project summary | The HITITA project', *The HITITA project* <http://blogs.uji.es/hitita/project-summary/> [accessed 18 October 2013].

[32] J. Ginther and A. Firey, 'T-PEN. Transcription for paleographical and editorial notation', *T-PEN*, 2012 <http://t-pen.org/TPEN/> [accessed 29 August 2012]; M. Terras, H. Cayless, and W. Noel, 'Introducing Tile 1.0', in *TILE | Text-Image Linking Environment* (College Park: Maryland Institute for Technology in the Humanities (MITH), 2011) <http://mith.umd.edu/tile/> [accessed 11 April 2012].

[33] P. A. Stokes, 'Describing Handwriting, Part I', *DigiPal - Digital Resource for Palaeography, Manuscripts and Diplomatic*, 27 July 2011 <http://digipal.eu/blogs/blog/describing-handwriting-part-i/> [accessed 12 October 2011] ; D. Stutzmann, 'Modélisation des signes graphiques (1)', *Paléographie médiévale*, 02 April 2012 <http://ephepaleographie.wordpress.com/2012/02/04/modelisation-des-signes-graphiques-1/> [accessed 30 August 2012].

[34] M. Rehbein, 'ESF Exploratory Workshop 'Digital Palaeography'. Würzburg, 20-22 July 2011', *Julius-Maximilians-Universität Würzburg. Zentrum für Digitale Edition*, 2011 <http://www.zde.uni-wuerzburg.de/veranstaltungen/digital_palaeography/> [accessed 8 September 2011].

[35] Ciula, 'Digital Palaeography'.

[36] Smith, 'Les formes de l'alphabet latin'.

ascenders of **b**, **h**, **k** and **l** behave likewise and they altogether are a criterion in some taxonomies,[37] but there also are scribes who provide some letters with loops and other not, or do not use the same writing angle for different letters: for example, ms. Paris, Bibl. nat. de France, latin 13722, written in the 1460s, uses a Semihybrida script with loops appearing only on **h** which is represented through two different allographs (the first one looped and having a straightforward, vertical descender, the other one written like an opened **b**).

The link to phonetics and linguistics could be revised as well, as stated above, on the hidden links between graphical evolution and phoneme frequencies. If we trace the systemic use of allographs, we may find new evidence for a distinctive implementation of formal differences and try to link them with phonetic processes and visual perception mechanisms. The scientific analysis of allographs based on this ontology should indeed remain very careful in its conclusions, since the script may have its own rules based on a graphical logic, as fully demonstrated by W. Oeser who also proved how almost unnoticeable letter-form variations were relevant to medieval scribes.[38]

Several disciplines are concerned with variability of scripts: epigraphy, diplomatics, palaeography, of course, but also linguistics, history, art history, communication studies, and even neurosciences. The graphical system is indeed a core concept which is relevant to every one of them. It may give new insights into otherwise unexplained ambiguities and phonetic evolutions, as well as reveal writing and reading strategies. Moreover, such an approach is clearly adequate for a computer-based analysis, which in turn may be the only way to finally address some core questions about the history of script. In this regard, some conditions are still not fulfilled.

2. ORIFLAMMS PROJECT: BUILDING A REFERENCE CORPUS FOR MEDIEVAL SCRIPTS

In order to address these challenges, that is, studying the linguistic and graphical system for Latin and vernacular scripts of the Middle Age, improving techniques of Computer Vision and Intelligent Writer Recognition, and building a formal ontology of forms, several research institutions came together. The research project ORIFLAMMS (Ontology Research, Image Feature, Letterform Analysis on Multilingual Medieval Scripts), funded for 3 years (2013-2016) by the French National Research Agency (*Agence Nationale de la Recherche*), aims at increasing knowledge of medieval scripts and multilingualism through a new, interdisciplinary approach; the teams will explore the graphical variability of scripts and match the results with a linguistics analysis (regularization of scripts and spelling).

2.1. CONCEPT

The ORIFLAMMS project brings together four public research units in the humanities and three research units in Engineering, Information Sciences and Technologies, including an industrial company : IRHT (Institut de Recherche et d'Histoire des Textes, leading); CESCM (Centre d'Études Supérieures de Civilisation Médiévale); ENC (École nationale des Chartes); ICAR (Interactions, Corpus, Apprentissage, Représentations) ; and in Computer Vision LIRIS (Laboratoire d'Informatique en Images et Systèmes d'Information); LIPADE (Laboratoire d'Informatique de PAris Descartes); A2iA.

As mentioned above, the lack of a large reference corpus is a major *desideratum*, which is why the teams will first gather and harmonize several research corpora, then increase and enhance them, in order to create a new reference corpus, covering the diversity of medieval scripts (handwriting to print, informal drafts to monumental inscriptions, from Carolingian times to the eve of Renaissance, from theology and liturgy to chancery rolls and accounts, both in Latin and vernacular languages, Old and Middle French, Middle English), and build a concordance of all written forms in the Middle Age, that is the required ontology.

---

[37] Derolez, *The Palaeography of Gothic Manuscript Books*.

[38] W. Oeser, 'Das "a" als Grundlage für Schriftvarianten in der gotischen Buchschrift', *Scriptorium*, 25 (1971), pp. 25-45, 1971; W. Oeser, 'Beobachtungen zur Strukturierung und Variantenbildung der Textura: Ein Beitrag zur Paläographie des Hoch- und Spätmittelalters', *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde*, 40 (1994), 359-439.

In order to create this concordance and move to a large-scale humanities computing project, ORIFLAMMS has to develop innovative image analysing tools and especially upgrade methods for aligning images to be able to precisely and accurately match words and letter-forms on available images. From then on, the research teams would be able to compute and measure the variability of scripts. The palaeographic and linguistic analysis will contribute with new knowledge about the evolution of writing in a multilingual environment like medieval society and, as such, enhance the comprehension of the scribal processes.

2.2. DATASETS IN THE REFERENCE CORPUS

The first step in the ORIFLAMMS project consists of the creation of a reference corpus, with images of scripts and graphically analyzed transcriptions from representative places and dates of medieval culture, and in interoperable formats. This reference corpus, which will be one of a kind due to its wide content, will be freely accessible, and give access not only to images, but also to graphically analysed transcriptions (allographetic transcriptions). The text will also be aligned with the image (with coordinates of pixels on the image). All data will be stored in an interoperable XML-TEI file for long term digital information preservation and access. The reference corpus has to be large and represent the productions of distinct regions, time periods, languages and scripts. It will rely on previous research and focus on available datasets and already digitised manuscripts:

- Pragmatic and diplomatic scripts from twelfth to fifteenth centuries will be represented by four groups, for example, Latin and French charters of Western France, Latin and French charters and cartularies of Burgundy, Registers of the French royal chancery, in Latin and French, produced in Paris in the 14th c., protocols and registers of notaries, in Latin, produced in Southern France in the 15th c.
- Book hands from the thirteenth to the fifteenth will be presented by six groups: French and Occitan datable or dated manuscripts; Latin datable or dated manuscripts; Latin works which circulated most widely in Europe and their translations in relation to the OPVS project,[39] allowing a comparative approach of the graphical systems in originals and in translations; bilingual manuscripts; *Queste du Graal* corpus, a single manuscript written in thirteenth century France;[40] *Projet Charrette* corpus;[41] *BFM – Manuscrits*.[42] As for the datable or dated manuscripts, the datasets are partly based on the catalogues of dated and datable manuscripts from 550 A.D. to 1600 A.D. preserved in France, established over more than fifty years of research,[43] and accompanied by a photographic collection of more than 9,800 images which can be used in the present project. Yet this corpus still has to be improved and enhanced in the nature, quantity, homogeneity and size of the digital images. At present, each manuscript is only represented by one or more 10 x 15 cm picture, which is barely enough for image characterization and features extraction; the pictures in the dataset are not homogeneous, given the great variety of layouts and the presence of flourished and illuminated initials or penwork; each image only partially covers the page (except for small size manuscripts); there are no transcriptions of the texts and limited metadata. The selection is based on the cross-matching of available digital reproductions and scientific analysis of the catalogues of dated manuscripts.
- Inscriptions, engraved and painted letters will be presented by three groups: Western France, 12th-15th c.; Burgundy, 12th-15th c.; Printed books, 15th c.

---

[39] G. Veysseyre, 'OPVS - Œuvres Pieuses Vernaculaires à Succès', 2011 <http://www.opvs.fr/> [accessed 21 July 2012].

[40] Marchello-Nizia, *Queste del saint Graal.*

[41] Pignatelli, 'L'archive du Projet Charrette'.

[42] A. Lavrentiev, 'BFM - Manuscrits', *Base de Français Médiéval — Old French Corpus*, 2009 <http://bfm.ens-lyon.fr/article.php3?id_article=177> [accessed 7 October 2011].

[43] Comité international de paléographie latine and D. Muzerelle, 'Manuscrits datés: État des publications', *Palaeographia* <http://www.palaeographia.org/cipl/cmd.htm> [accessed 12 May 2012]; D. Muzerelle, 'Manuscrits datés (France) - Index général interactif', *Aedilis*, 2006 <http://aedilis.irht.cnrs.fr/cmdf/> [accessed XXX].

This Reference Corpus may increase in the future in order to give not only a training set, but also a validation set of images and texts.

2.3. IMAGES AND METADATA IN THE REFERENCE CORPUS

According to the project aims, the reference corpus has to assemble images, texts and metadata. The image specifications follow the common standards: captured in colour, 300 dpi minimum, and stored in a format suitable for long term preservation, such as TIFF. Some experiments are planned with lower quality images, especially digitized microfilms, to compare the results and determine if the very large collections of microfilms worldwide could be useful for this research in the future.

The textual data presented by the images will be presented in XML-format allowing the transcription to be fully aligned and, eventually, to store the different readings of a subset of signs (e.g. abbreviated words, useful for allographetical research and the formation of abbreviation lists by frequency, and plain words, necessary for checking the forms in dictionaries). This format has to be compliant with the TEI-P5 Guidelines:[44] it is foreseen that the textual part of the corpus can be enhanced and that further metadata can be added, such as a semantic or morphosyntactic analysis and the TEI-compliant encoding provides this ability as well for the alignment metadata (through the facsimile-module) as for the semantic and linguistic information. A work package within the project deals with the specifications of a TEI compliant schema, the choice between the several available solutions of the TEI Guidelines, and the creation and implementation of new entities. At the end of the project, graphical forms, plain text and allographetic transcription should be available in the reference corpus, and could be presented in an online interface as they are on the *Queste du Graal* website.

Plain text:

Et quant il sont a la porte si apele li escuiers, et l'en
li huevre, et il descendent, et entrent enz, et quant
cil de laienz sorent que Lancelot estoit venuz si li vont

Allographetic transcription:

Et quant il ſont a lapoꝛte ſi apele li eſcuíerſ . ꝛ len
li hueure . et il deſcendent . ꝛ entrent enz . ꝛ qn̊t
cil de laienz ſoꝛent que lanc̄. eſtoit uenuz ſi li uo̅t

Source: *Queste del saint Graal Édition numérique interactive du manuscrit de Lyon (Bibliothèque municipale, P.A. 77)*, ed. by C. Marchello-Nizia and A. Lavrentiev : ms K, folio 160, recto, col. b (http://portal.textometrie.org/bfm/?command=documentation&path=/GRAAL)

Figure 1: Allographetic transcription

The length of each sample in the reference corpus will vary from one document to the other. Since charters and inscriptions are shorter than books, they will be presented as complete entities. It will be determined during the project how long the allographetic transcriptions should be for book handwriting. Some of the datasets are already available (*Queste du Graal*, *Charrette*, *BFM-Manuscripts*, Burgundian charters), but are heterogeneous, due to the nature of each text (differences in length from one charter to the other or between several literary works) or to the encoding choices (record type editions, diplomatic editions, normalizing editions). If we may unify the encoding to a certain extent during the project (e.g. conversion from TEI-P4 to TEI-P5[45]), it is not possible to homogenize the whole range of medieval artefacts. The length of the samples will depend of the typology of documents: charters and

[44] TEI Consortium, 'TEI P5: Guidelines for Electronic Text Encoding and Interchange', 2013-2007 <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/> [accessed 2 November 2013].

[45] TEI Consortium, 'TEI P5'.

inscriptions are shorter texts and should obviously be presented in their full length, which is a solution that scholarly editions have already adopted. As for book scripts, one may dream of a synoptic transcription of all manuscripts of one work, but it is obvious that a medieval bestseller such as the *Roman de la Rose* with no less than 80 digitized manuscripts available to the project cannot be transcribed and studied in full. Instead of setting a standard length for book handwritings at the beginning of the project, a pragmatic approach has been preferred: using the available material as being a more or less conscious statement about scientific requirements in the humanities and enhancing it as needed by computer vision tools. In former research on abbreviations and graphical systems, the average length is of 700 words (*BFM-Manuscrits*) but some inquiries were made on Psalm 101, which is 120 words in length.[46] A reasonable training set for some of the tools used by the project partners has to comprise at least 1000 words, or more according to the variability of the script. In turn, as this project is focused on measuring variability, the required length is part of the measurement and this length may be slightly different from one manuscript to one other until the quality of the results ceases to improve.

The reference corpus will be completed with descriptive metadata, especially for support, layout, date and place of production, but also on author, title and text typology, so that a cross-analysis can be provided. The most convenient solution is to store the information directly within the transcription document, in the following elements:

- msDesc>physdesc>objectDesc>supportDesc>support
- msDesc>physdesc>objectDesc>layoutDesc>layout @columns @writtenLines
- msDesc>history>origin>origDate
- msDesc>history>origin>origPlace

Some additional remarks about the palaeographical evidence and preliminary statements on the script typology according to diverse extant taxonomies would be stored in msDesc>physdesc>scriptDesc. The msDesc element is located as teiHeader>source>listWit>witness>msDesc. According to the nature of the corpus, however, the number, length or tag sets of related witnesses and transcriptions may vary.

2.4. ALIGNING THE TEXT AND THE IMAGE

The reference corpus will be enhanced by the alignment between text and image. In the present project, the goal is to achieve a letter-by-letter alignment for set hands, by which each unit of the transcribed text could be associated to its graphical representation. Given the amount of textual data in the corpus, an alignment done by hand, by selecting the letters one-by-one, as in the DAmalS and DigiPal projects, is not an option.[47] This can only be the result of an automated or semi-automated process. The development of an alignment tool is also an integral part of the project and will help palaeography move to a more industrial approach within the large field of digital humanities and get over its artisanal and paradoxical stage, in which the results of a completely automated and high level image analysis has to be validated against a very traditional expertise, largely based on a letter-by-letter observation.

As the transcriptions are TEI-P5 compliant, this format will also be adopted for storing the alignment data.[48] The main reason for this choice over various metadata standards which could be used, such as ALTO (Analyzed Layout and Text Object), METS (Metadata Encoding Transmission Standard), is that it is the only one that allows a complete analytical process from the text and its alphabetical representation to image analysis.[49] This format also

---

[46] Lavrentiev, 'Pour une méthodologie'; Lavrentiev, 'Typologie textuelle'; C. Bozzolo, D. Coq, D. Muzerelle, and E. Ornato, 'Les abréviations dans les livres liturgiques du XV^e^ siècle : pratique et théorie', in *La face cachée du livre médiéval : l'histoire du livre vue par Ezio Ornato, ses amis et ses collègues*, ed. by E. Ornato (Roma: Viella, 1997), pp. 555-65.

[47] Hofmeister and others, 'Forschung am Rande'; Stokes and others, *DigiPal*.

[48] TEI Consortium, 'TEI-P5'. Consortium Oriflamms, "Spécification du format XML-TEI pour l'alignement texte-image. 1. Structure et convention de nommage", *Écriture médiévale & numérique* (11 September 2016) <http://oriflamms.hypotheses.org/1442>; Consortium Oriflamms, "Spécification du format XML-TEI pour l'alignement texte-image. 2. Bonnes pratiques d'encodage", *Écriture médiévale & numérique* (12 September 2016) <http://oriflamms.hypotheses.org/1510>.

[49] Library of Congress, 'ALTO: Technical Metadata for Optical Character Recognition (Standards, Library of Congress)', 2011 <http://www.loc.gov/standards/alto/> [accessed 30 April 2011]; Library of Congress,

allows semantic enhancement with mark-up elements such as place and person names or syntax, whose influence on the scribal performance are within the scope of this project.

Structure and layout extraction is the first part (columns and lines, such as in T-PEN).[50] As for existing transcriptions and editions, such a tool is not enough, unless one is ready to copy-paste each line. As a consequence, there is a need for a new tool which embeds OCR-functionalities, linguistic dictionaries with lemmatisation features (Latin and French), as well as a dictionary of abbreviations, in order to automatically proceed to the alignment and also to detect letters or words that cannot be aligned. Extracting the structure first, then aligning the lines, the words and, only at the end, the letters, should streamline the process, since, at each stage, the human eye can perform very quickly the necessary verification and validation if it is provided with an adequate monitoring tool. At a letter level, the alignment process has to be supported by a monitoring tool, allowing easy verification of all characters. Such tools already exist, generally in relation to OCR or word spotting software. A major function which will be implemented once the tool is developed is a categorization tool, since each recognised letter can be assigned to another or a new sign, with consequences for the transcription itself.

This step is perhaps the more challenging technological issue for the partners in this project since any further palaeographical research will rely on a fully analysed corpus. A prerequisite is the identification of classes of handwriting or scripts that are not accessible to the OCR tools and the gradual extension from simpler scripts towards more challenging ones. This alignment tool is the technological core of the project because it is an integral part of the computer assisted transcription and learning system.

2.5. BUILDING THE ONTOLOGY AND ENHANCING THE CORPUS

The alignment software will produce a powerful data set for new research which includes *de facto* a complete dictionary of glyphs and forms that ought to be exploited in parallel and in a cross-fertilisation process, as well for new technological developments towards a better understanding of the historical evolution of scripts.

Each letter-form category, roughly corresponding to an allograph, will be created first according to the palaeographical state of the art (that is based both on the formal evidence and on the dynamical reality of the *ductus*), then has to be envisioned and measured 'objectively', that is with similarity computing. It may well be that some letter-forms that seem very close to palaeographers will be separated by the machine. Even if the system will initially be fed with existing allographetic transcriptions, this ontology and the monitoring tool intend to ease and modify the transcription mechanism. This is because currently the text is transcribed first (either directly, or taken from a critical edition) then re-read and processed in order to add abbreviations and allographetic information, based on which allographs have been chosen within the framework of the project.[51] In this project, the allographetic transcriptions are meant to inform the OCR functions, but, once alignment is achieved, the allographetic data is added directly at the letter level, by selecting all similar forms and assigning them to a single class. This stage is crucial because it means not only classifying the forms and confronting two very different perceptions of the graphical evidence, but also suggesting a heuristic tool for script analysis. It will probably be elaborated in several steps and by testing several hypotheses and analysing the results according to the diverse extant palaeographical theories.

Through this project, it is also intended to contribute to the two major achievements towards interoperability and exploiting allographs, which are the TEI gaiji module and the MUFI extension to Unicode (Medieval Unicode Font Initiative).[52] Both have been merged into gBank thanks to the Manuscriptorium project, but the accumulation of

---

'Metadata Encoding and Transmission Standard (METS) Official Web Site', 2011 <http://www.loc.gov/standards/mets/> [accessed 30 April 2011].

[50] Ginther and Firey, 'T-PEN'.

[51] D. Stutzmann, 'Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer and aller plus loin ?', in *Kodikologie und Paläographie im digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2*, ed. by F. Fischer and others (Norderstedt: Books on Demand, 2010), pp. 247-77.

[52] TEI Consortium, 'TEI-P5'; Medieval Unicode Font Initiative, 'MUFI character recommendations', 2009 <http://www.mufi.info/specs/> [accessed 12 April 2011].

ambiguous solutions, the lack of glyphs, and the lack of associated ergonomic tools make them difficult to use.[53] Those tools and standards need to be regularized and expanded. Such an improvement can only be achieved through a complete analysis of a large corpus, such as the reference corpus described above.

2.6. IMAGE ANALYSIS, CLUSTERING AND MEASURING THE VARIABILITY

During the project, image analysis techniques will obviously be used in due course. As already mentioned, clustering functionalities are used to facilitate the time consuming task of validating the alignment or to improve the alignment itself, and will also serve to provide a new perspective while extending the ontology.

At the first level, research on the variability of scripts can be performed with computational and statistical models on the letter-form (intra-allographic, inter-allographic and inter-letter analysis), which is a step beyond the study of the graphical system and its consistency or variability. Considering the graphical components is an integral part of the study of medieval scripts, since physical factors can explain complex historical evolutions (friction of the nib on the support and the impossibility of drawing some strokes with a bevelled nib). This research is similar to variance analysis in linguistics.

At a second level, research in computer vision should be performed. The ORIFLAMMS project can rely on the results of the research program GRAPHEM (Grapheme based Retrieval and Analysis for PalaeograpHic Expertise of medieval Manuscripts, 2008-2011), funded by the French National Research Agency, which aimed at improving data mining and image processing techniques applied to medieval scripts and their classification with several methods.[54] Yet global approaches make it very difficult to relate the results back to the visual differences that the palaeographer can observe. Local methods using features which in the first instance sound familiar to the methods of palaeographers, such as curvature, thickness, roundness, angularity, pen angle, etc, (e.g. Freeman chain code which consists in the analysis of strokes direction and angularity based on the contours) give results that cannot be traced back to the morphological features that the palaeographers are used to observing. Through the multidimensionality of all the processes, the results are reduced and projected to common axes and are loosened up from a one-to-one relationship with palaeographical features. The codebook method developed by H. Daher is very promising in this regard and could be implemented for a letter-level analysis, since it is based on a significant and understandable palaeographical component analysis (classification of strokes), but goes further than the 'palaeographer's eye' could ever imagine[55].

The GRAPHEM project will be developed towards a graphical analysis of variance. New features have to be developed and studied in ORIFLAMMS; these features have to be theoretically independent of alphabet and language in order to use OCR and IWR (Intelligent Word Recognition) calculations and combine global and structural features, as a means to measure the legibility of texts. The idea is to benefit from the pattern similarity measure, validating using a modelled ideal letter-form as validation, to study the variability itself: how far and how dispersed are the measured forms from the ideal type? Both partners LIRIS and LIPADE have demonstrated experience in the domain of legibility, word spotting, and legibility measurement for handwriting with statistics, fractals, Gabor, Hermite transforms, curvelets, skeletonization and multi-oriented text lines detection.[56] A2iA will

---

[53] Oxford University Computing Services, 'ENRICH gBank viewer - v.1.00', 2009 <http://www.manuscriptorium.com/apps/gbank/gbank_table.php> [accessed 30 April 2011]; Stutzmann, 'Paléographie statistique'; T. A. McAllister, 'Computing with Medieval Characters: Updates to MUFI and Unicode', *Bulletin of International Medieval Research*, 12 (2006), pp. 40-7.

[54] LIRIS, 'GRAPHEM : Projet ANR' <http://liris.cnrs.fr/graphem/> [accessed 1 May 2011]; F. Cloppet and others, 'New Tools for Exploring, Analysing and Categorising Medieval Scripts', *Digital Medievalist*, 7 (2011) <http://digitalmedievalist.org/journal/7/cloppet/>; D. Muzerelle, 'À la recherche d'algorithmes experts en écritures médiévales', *Gazette du livre medieval*, 56–57 (2011), pp. 5-20.

[55] H. Daher and others, 'Étude de la dynamique des écritures médiévales. Analyse et classification des formes écrites', *Gazette du livre médiéval*, 56–57 (2011), pp. 21-41.

[56] V. Eglin, 'Contribution à la structuration fonctionnelle des documents imprimés: Exploitation de la dynamique du regard dans le repérage de l'information' (unpubl. doctoral thesis, Institut national des sciences appliquées de Lyon, Villeurbanne, 1998); V. Bouletreau and others, 'Synthetic Parameters for Handwriting Classification', in

train hybrid recognizers based on neural networks and Hidden Markov Models on the corpora produced by the project.[57] The partners intend to develop the combed approach of letter and word-level analysis while creating a large dataset. The partners will follow the suggestion made by H. Bunke's team, which recommended associating letter-level segmentation with text recognition systems in order to train the systems at a letter and at the word levels, but also T. Rath's work on ASCII tagging and the creation of a training dataset for text recognition and content-based image retrieval.[58]

The ORIFLAMMS project intends to proceed to pattern similarity calculations, not only within each letter class to distinguish between allographs, but also a transverse similarity check in order to identify ambiguous forms and deepen the analysis of their uses and correlations. We might indeed imagine a rich system for computer-aided transcriptions which would contain not only a complete set of alphabetical forms belonging to the Middle Ages (or several sets to suit the different scripts), but also specific rules that help to distinguish ambiguous letter-forms from one another. As such, there should be a reciprocal improvement of palaeographic expertise and computer engineering, in categorising and characterising the morphologies and exploring a way to disambiguate complex nodes in the formal ontology.

3. CONCLUSION

Variability is a key concept for the humanities and is also a core question and challenge for Computer Science because of the difficulties it presents for any attempt to automate the analysis. In the newly opened dialogue between humanities and computer science, variability may well be the subject which needs to be addressed first and foremost,

---

Proceedings of the 4th International Conference on Document Analysis and Recognition, 2 vols, (Los Alamitos, CA: IEEE Computer Society, 1997), I, 102-6; V. Eglin, 'Approches perceptives et cognitives en analyse automatique d'images de documents', *TSI : Technique et science informatiques*, 25 (2005), pp. 523-51;. Eglin and A. El Abaed, 'Frequencies Decomposition and Partial Similarities Retrieval for Patrimonial Handwriting Documents Compression', in *Proceedings of the 8th International Conference on Document Analysis and Recognition*, 2 vols (Los Alamitos: IEEE Computer Society, 2005), II, 996-1000; V. Eglin, S. Bres, and C. Rivero, 'Hermite and Gabor Transforms for Noise Reduction and Handwriting Classification in Ancient Manuscripts', *International Journal on Document Analysis and Recognition*, 9 (2007), pp. 101-22; S. Bres, V. Eglin, and C. Volpilhac-Auger, 'Evaluation of Handwriting Similarities Using Hermite Transform', in *Tenth International Workshop on Frontiers in Handwriting Recognition*, ed. by G. Lorette, H. Bunke and L. Schomaker (La Baule: Suvisoft, 2006); G. Joutel and others, 'Curvelets Based Feature Extraction of Handwritten Shapes for Ancient Manuscripts Classification', in *Document Recognition and Retrieval XIV. 30 January-1 February 2007, San Jose, California, USA*, ed. by X. Lin and B. A. Yanikoglu (Bellingham: SPIE-IS&T, 2007), pp. 65000D1-12 <http://dx.doi.org/10.1117/12.704450>; G. Joutel, 'Analyse multirésolution des images de documents manuscrits : Application à l'analyse de l'écriture' (unpublished doctoral dissertation, Institut National de Sciences Appliquées de Lyon, Villeurbanne, 2009); Y. Leydier, F. Lebourgeois, and H. Emptoz, 'Outils informatiques pour l'indexation et la recherche dans les manuscrits médiévaux', *Le Médiéviste et l'Ordinateur*, 45 (2006) <http://lemo.irht.cnrs.fr/45/indexation.htm>; Y. Leydier, J. Duong, and A. Ouji, *Ulysse 0.3g09*. 2006.; H. Daher and others, 'Décomposition des manuscrits anciens en graphèmes et construction des codes book basée sur la coloration de graphe', in *COmpression et REprésentation des Signaux Audiovisuels (CORESA)* (Lyon: Coresa, 2010), p. 105; H. Daher and others, 'Ancient Handwritings Decomposition into Graphemes and Codebook Generation Based on Graph Coloring', in *International Workshops on Frontiers in Handwriting Recognition (ICFHR)*, (Los Alamitos: IEEE Computer Society, 2010), pp. 119-24; H. Daher and others, 'A New Approach for Centerline Extraction in Handwritten Strokes: An Application to the Constitution of a Code Book', in *Document Analysis Systems*, XXX (2010), pp. 425-32.

[57] T. Bluche and others, 'The A2iA Handwritten Arabic Text Recognition System at the OpenHaRT2013 Evaluation Campaign', *Proceedings of the 11th IAPR International Workshop on Document Analysis Systems [DAS 2014]*, (Los Alamitos: IEEE, 2013), pp. 161-5.

[58] E. Indermühle, M. Liwicki, and H. Bunke, 'Combining Alignment Results for Historical Handwritten Document Analysis', in *Proceedings of the 10th International Conference on Document Analysis and Recognition* (Los Alamitos: IEEE Computer Society, 2009), pp. 1186-90; A. Fischer and others, 'Transcription Alignment of Latin Manuscripts using Hidden Markov Models', in *Proceedings of the 1st International Workshop on Historical Document Imaging and Processing (HIP)* (New York, NY: Association of Computing Machinery, 2011), pp. 29-36; J. L. Rothfeder, R. Manmatha, and T. M. Rath, 'Aligning Transcripts to Automatically Segmented Handwritten Manuscripts', in *Proceedings of the 7th International Workshop on Document Analysis Systems*, ed. by Horst Bunke and A. Lawrence Spitz (Berlin: Springer, 2006), pp. 84-95.

because variability is an underlying, constant reality for some and a significant obstacle for others. Identifying peaceful islands of stability within an evolving world of languages and script would allow some progress in the much debated question of taxonomy of Gothic scripts and in automated handwriting analysis.

Aware of the scientific, technologic, industrial and societal issues in order to analyse the evolution of writing systems and graphical forms during a long period (the Middle Ages) and according to their production contexts (informal, documentary, book scripts) and languages (Latin or vernacular), the partners of the ORIFLAMMS research project gathered to establish an ontology of forms, analyse the graphical structures of scripts and upgrade a linear, textual approach to a visual, two- or three-dimensional one, in order to create new knowledge for linguistics and cultural history (palaeography, epigraphy, diplomatics).

The challenges are many: not only to gain a new understanding of what script is (the writing process and formal issues), but also to change the conditions of research on scripts; to establish a reference corpus for palaeographers and linguists, i.e. for any research on the history of the languages and their written form; to create new exploitation tools in an open-source software suite; to implement standards and good practices and diffusing them through the developed tools; to create an ontology of signs; to build an interoperable and extendable platform, meeting the needs of long-term preservation. These steps lead to overcome the usual borders between related sciences (epigraphy, palaeography, linguistics) and between human and computer sciences. Indeed, the excessive compartmentalisation in the human sciences, according to the support or context (epigraphy, book script, diplomatic script, printed script) or the language (Latin and vernacular) is a structural obstacle, as is also the misconception of palaeography as an 'auxiliary science' and not as an integral part of cultural and intellectual history which has lead to a complete ignorance of reciprocal influences of forms, alphabetical chains, and graphical systems in the historical development of scripts and languages. In the collaboration between human and computer sciences, the reference corpus and the associated tools will resolve the difficulty of encoding new sources and exploiting diverse and dispersed corpora and, especially, apprehending their internal variability.

Building and analysing the corpus can only be the result of an intensive cooperation between human sciences (linguists, palaeographers and epigraphists) and computer sciences, taking into account the needs of analysis and the issues of ergonomics and usability. The innovative tool is not only about producing new encoded data; it is also part of their exploitation. ORIFLAMMS plans to offer a new method for the study of scripts by analyzing graphical variability. The latter will be considered through image analysis on a two-dimensional level, and through computational linguistics for the variability of morphosyntactic and graphical codes in Latin and the vernacular. The open source software of computational linguistics (TXM) should indeed be upgraded and documented as part of this project.

The strategy described above has been implemented, even if minor changes intervened during the time frame of the project. As for the aims described in this article, it is worth recapitulating some of the publications and realisations brought up by the partners.

As for the Reference Corpus, which should allow a joint analysis on different types of sources, it is published on GitHub <https://github.com/oriflamms> and encompasses charters, manuscripts and diplomatic material. The full range of the medieval diversity is not gathered online at this place, since the epigraphic material will be added in 2018. The intended corpus of 400 acts from French royal chancery registers was used for the European research project HIMANIS in which some of the partners of ORIFLAMMS have indexed the full text of some 200 medieval registers, i.e. 70'000 pages. As a result, we decided not to publish the small corpus of 400 acts, but intend to release a much larger, annotated corpus. The full text index can be searched online <http://prhlt-kws.prhlt.upv.es/himanis/>. For the sake of interoperability and standardisation, we published a format and best practices, that is a completely TEI compliant format with requirements to allow for a steady use for palaeographical studies[59]. It is based on a stand-off encoding of texts, zones and annotations.

[59] Consortium Oriflamms, "Spécification du format XML-TEI pour l'alignement texte-image. 1. Structure et convention de nommage", *Écriture médiévale & numérique* (11 September 2016) <http://oriflamms.hypotheses.org/1442>; Consortium Oriflamms, "Spécification du format XML-TEI pour

The alignment software is published as open source software <https://github.com/Liris-Pleiad/oriflamms>. This software also gives a tabular view of words and letters, it allows the monitoring and validation of the alignment and integrates an allograph classification tool. The results of the alignment can also be search in an online interface at <http://oriflamms.teklia.com/>. The open source and modular platform TXM for lexicometry and text statistical analysis, is enhanced with functionalities to perform queries on palaeographical features (letter-forms and abbreviations). Beyond their own work on script classification, the Oriflamms partners organized two competitions on script classifications,[60] and fostered the cross-disciplinary reflection[61].

---

l'alignement texte-image. 2. Bonnes pratiques d'encodage", *Écriture médiévale & numérique* (12 September 2016) <http://oriflamms.hypotheses.org/1510>.

[60] Dominique Stutzmann, *Competition on the Classification of Medieval Handwritings in Latin Script* (Paris: Institut de Recherche et d'Histoire des Textes, 2016) <http://clamm.irht.cnrs.fr/>; F. Cloppet, V. Eglin, V. Kieu, D. Stutzmann, and N. Vincent, 'ICFHR2016 Competition on the Classification of Medieval Handwritings in Latin Script', in *Proceedings of the International Conference in Frontiers on Handwriting Recognition* (Shenzhen, China, 2016), pp. 590–95 <http://ieeexplore.ieee.org/document/7814129/>; M. Kestemont and D. Stutzmann, 'Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts', *Speculum*, 92 (2017) <http://www.journals.uchicago.edu/doi/10.1086/694112>.

[61] D. Stutzmann and S. Tarte, 'Digital Palaeography: New Machines and Old Texts : Executive Summary', ed. T. Hassner, D. Stutzmann, and S. Tarte, Dagstuhl Reports 4, no. 7 (2014): 112–134 (112–114, 132), <https://doi.org/10.4230/DagRep.4.7.112>; D. Stutzmann and S. Tarte, 'Paléographie Numérique / Digital Palaeography | Fondation des Treilles' (Tourtour, Fondation des Treilles, 2017) <http://www.les-treilles.com/paleographie-numerique-digital-palaeography/> [accessed 11 March 2017].