“ Modéliser le continuum latino-roman aux alentours de l’an 800 : de la sociolinguistique à l’intelligence artificielle ”

Florian Cafiero, Rémy Verdo

VERDO (Rémy) et CAFIERO (Florian), « Modéliser le continuum latino-roman aux alentours de l'an 800 : de la sociolinguistique à l'intelligence artificielle », dans Latin vulgaire – latin tardif : actes du XIII[e] colloque international sur le latin vulgaire et tardif (Budapest, 03-07 septembre 2018), p. 453-466.

[p. 453] FLORIAN CAFIERO – REMY VERDO

MODELISER LE *CONTINUUM* LATINO-ROMAN AUX ALENTOURS DE L'AN 800: DE LA SOCIOLINGUISTIQUE A L'INTELLIGENCE ARTIFICIELLE*

Summary: The shifting from a "dilated diasystem", when Latin becomes more and more complex, to two distinct linguistic systems, has already been modelled (e.g. by Pulgram, 1950; Berschin, 1986). Relying on their authors' extensive experience, these models however leave some problems unaddressed. In particular, they consider the language of a specific period as a homogenous whole. Thus, they mostly ignore the variations of registers existing in the language at a same time, sometimes in the same text. In this paper, we propose a method to systematically study the evolution of the various registers used in texts written in Carolingian ages, with regards to later Merovingian ones. Some of the results can be obtained through computerized statistical analysis implementing some artificial intelligence: the tagging of whole sentences can be applied to a large amount of texts, too large to be analyzed otherwise. Using an annotated corpus as training data, we develop an artificial intelligence that identifies the various registers used in a text. We intend to implement it on a large selection of texts written during the same period.

Key words: diachronic sociolinguistics, text mining, Carolingian Latin, digital humanities, statistics, early Romance

1. Introduction

Cette présentation porte sur une période centrée sur la toute fin du VIII[e] s., qui intervient un demi-siècle après le basculement typologique de la langue latine vers la langue romane dans la future zone de langue d'oïl[1], tel qu'il s'observe dans les textes.

Le corpus textuel étudié ici mêle deux genres très complémentaires, quoique traditionnellement étudiés de manière compartimentée: des vies de saints, et des chartes. La méthode d'analyse mise en œuvre sur ce corpus dans le cadre d'une thèse [p. 454] de l'École des chartes[2] correspond à une sorte d'archéologie du langage, pratiquée "de tête".

Dans les pages qui suivent, nous exposerons le fonctionnement de cette première démarche, puis nous examinerons les traits fondamentaux de ses résultats, pour enfin montrer la manière dont ces derniers ont été en partie reproduits à partir d'une méthode de fouille de textes, automatisée et donc capable de préparer l'étude de corpus bien plus larges.

2. Modéliser les registres de langue: le cadre de départ

À partir du milieu du VII[e] s., on voit apparaître peu à peu, dans les textes, des énoncés longs (propositions avec verbe conjugué, et plus rarement des phrases entières) dont la structure morphologique, lexicale et topologique contient peu d'éléments, voire aucun élément propre au latin et n'existant donc pas en langue romane. Il s'agit par exemple

[1] BANNIARD, M.: *Viva voce: communication écrite et communication orale du IV[e] au IX[e] siècle en Occident latin.* Paris 1992, 488–490.

[2] VERDO, R.: *La reconfiguration du latin mérovingien sous les Carolingiens: étude sociolinguistique des diplômes royaux et des réécritures hagiographiques (VII[e]-IX[e] siècle).* Thèse de l'École des chartes. Paris 2010.

d'un verbe comme *rogaret*, forme d'imparfait du subjonctif inexistante dans les langues romanes où lui succède la forme *roga(ui)sset*[3].

Un siècle plus tard, vers 750, la langue latine d'usage courant a achevé de basculer d'un état "tardif" à un état si "innovant" (protofrançais) que les locuteurs dépourvus de culture grammaticale ne comprennent plus l'essentiel des discours construits avec une certaine fréquence de structures proprement latines. Par exemple, un bref énoncé tel que *in conpensatione huius rei*, lu dans un précepte de Louis le Pieux de 821[4], est un exemple de syntagme "médian" destiné à éviter l'emploi de la langue romane autant que celui d'une langue latine trop archaïque. Il suffit, pour s'en convaincre, de réécrire la tournure en latin de plus en plus ancien et en latin de plus en plus romanisant:

[p. 455]

Dans ce schéma, on voit que le message énoncé est susceptible de variations stylistiques dans un *continuum* linguistique qui mène insensiblement d'un système langagier (le latin) à un autre (la langue romane, dans sa réalisation française). Ces variations sont employées différemment selon les époques, les situations d'énonciation, les zones d'émergence des futurs dialectes, et selon le niveau culturel du locuteur.

L'on voit, par cet exemple, que l'opposition entre le "classique" et le "vulgaire", autrement dit entre le latin traditionnel et l'ancien français, matérialisée par le positionnement extrême de ces derniers dans ce diasystème, et appuyé par le contraste entre le bleu (couleur froide, pour la tradition) et le jaune (couleur chaude, pour l'innovation), ne constitue que les contours du modèle, les limites du diasystème. C'est effectivement le *continuum* séparant ces deux pôles qui est essentiel pour comprendre l'histoire de la langue, et pour lequel il manque à la fois des descriptions fines et un outillage conceptuel.

---

[3] Sur cette forme, les rares exceptions sont concentrées dans cette zone de latinophonie ultime que semble bien avoir été la Sardaigne: on y trouverait encore la forme dans certains dialectes, comme celui de Logudoro: TOGEBY, K.: Le sort du plus-que-parfait latin dans les langues romanes. *Cahiers Ferdinand de Saussure* 23 (1966) 183.

[4] Précepte de Louis le Pieux, 6 novembre 821, Thionville (Archives nationales, sous-série K8, nº 11, in *Diplomata Karolinorum: recueil de reproductions en fac-similé des actes originaux des souverains carolingiens conservés dans les archives et bibliothèques de France*. Éd. LOT F., LAUER P. et TESSIER G. Toulouse–Paris 1936 [sans transcr.], t. 2, pl. XI).

Pour l'époque carolingienne, une modélisation complète du *continuum* a été proposée pour la première fois par Michel Banniard en 2008[5]. Fondée sur l'observation de textes de genres très variés, elle dégage cinq registres de langue. L'usage de ces registres est déterminé par la portée pragmatique des textes: ainsi, le registre 5 correspond aux textes de portée générale, adressés à une large partie du peuple, et le registre 1 correspond aux textes dont la portée est restreinte aux auditeurs munis d'un solide savoir grammatical:

> Registre 5. – Protofrançais direct: commandements à l'intérieur du palais adressés aux domestiques, esclaves, etc. Oralité immédiate en accent local. Sous le terme protofrançais, on comprendra toutes les variétés dialectales dont les contours sont en voie d'émergence (lorrain, champe[p. 456]nois, wallon…). Emploi évidemment massif. Coïncidence profonde avec la parole ordinaire relâchée (même si la graphie masque la prononciation).
>
> Registre 4. – Latin à phrasé[6] protofrançais saupoudré de quelques latinismes aléatoires: commandements lors de cérémonies solennelles collectives; rapports oraux de missions sur l'état d'abbayes, de corps d'armée, certains polyptyques, etc. Oralité démarquée en diction plus soignée, mais en accent également roman. Emploi ouvert à de vastes pans des activités juridico-notariales; masque mince de *grammatica* (même si l'orthographe est impeccable). Réalisation orale relâchée en phonétique quotidienne.
>
> Registre 3. – Latin à phrasé protofrançais combiné à des séquences plus franchement latines, sorte de *lingua mixta*: rapports écrits de mission des *missi dominici*; capitulaires, notamment le *De uillis*; serments. Employé massivement par les élites carolingiennes, pratiquant une *mimésis* limitée des registres 1 et 2, sans admettre complètement les registres 4 et 5. Réalisation orale polie limitant les compromis avec la phonétique naturelle.
>
> Registre 2. – Latin en *stylus simplex* comprenant des séquences de protofrançais masqué: préambules des capitulaires; corps des lettres dans les correspondances; traités particuliers d'éducation. Partagé par une élite plus étendue, juristes, chanceliers royaux, certains évêques et abbés. Réalisation orale soutenue, correspondant à une certaine *distinctio*.

---

[5] BANNIARD, M.: Du latin des illettrés au roman des lettrés: la question des niveaux de langue en France (VIII[e]-XII[e] siècle). In VON MOOS, P. (éd.): *Zwischen Babel und Pfingsten: Sprachdifferenzen und Gesprächsverständigung in der Vormoderne (9.-16. Jh.) Akten der 3. deutsch-französischen Tagung des Arbeitskreises "Gesellschaft und individuelle Kommunikation in der Vormoderne" (GIK) in Verbindung mit dem Historischen Seminar der Univ. Luzern. Höhnscheid (Kassel), 2006.* Münster 2008, 269–286. Le modèle a été légèrement affiné dix ans plus tard dans BANNIARD, M.: Comment le latin parlé classique est devenu le français parlé archaïque: pour une historicisation et une modélisation innovantes (bréviaire). In CARLIER, A. – GUILLOT-BARBANCE, C. (éds): *Latin tardif, français ancien: continuités et ruptures.* Berlin–Boston 2018, 26–27.

[6] Cette notion de phrasé, chère à M. Banniard, attend encore un article définitoire. On peut considérer qu'il s'applique à des énoncés longs comportant plus qu'un seul syntagme. Il s'agit donc au minimum d'une proposition avec un verbe conjugué, voire d'une ou plusieurs phrases entières. La détermination du phrasé repose sur l'observation de la manière dont sont choisis et surtout disposés les éléments au sein de ces énoncés. C'est le caractère innovant, neutre ou archaïsant des structures (vocabulaire, morphologie, syntaxe, mais encore et surtout l'ordre des mots et des syntagmes, ainsi que la qualité et la quantité des disjonctions) qui est à quantifier pour déterminer si l'énoncé est finalement plus roman que latin. Dans un tel cas, si un énoncé théoriquement latin contient quelques formes archaïques dans une disposition résolument innovante, on peut dire que cet énoncé est en phrasé protoroman.

> Registre 1. – Latin en *sermo altus* ne comprenant que des séquences brèves de type roman: vies de saints réécrites; traités de théologie et de controverse doctrinale (*Libri carolini*); poésies soit de forme classique, soit rythmiques. Réservé au premier cercle des *grammatici*. Ultra-minoritaire. Réalisation orale soignée tentant de restaurer une syllabation complète.

Cette modélisation est commode: elle a l'avantage de donner du sens à la variation stylistique que l'historiographie a enfin cessé d'interpréter comme une inconstance dépendant du niveau de connaissances grammaticales de chaque auteur. Le modèle se structure principalement en fonction de la fréquence des formes et tournures anciennes ou neuves, et fait place secondairement à la complexité de la macrostructure (essentiellement: la longueur des phrases et leur degré d'hypotaxe). C'est cette question de macrostructure qui constitue la principale faiblesse de ce modèle carolingien qui essaie (heureusement, avec modération) de combiner deux approches qui sont parfois contradictoires, déjà à cette époque, et le seront de plus en plus nettement avec l'accès de la langue romane à un statut littéraire. En effet, comme l'expose le [p. 457] titre de la publication de 2008 où M. Banniard présente pour la première fois son modèle,

> – on peut concevoir un texte en langue typologiquement protoromane et d'un haut niveau rhétorique (usage de périodes amples, dont la structuration hypotactique se répartit en plus de deux niveaux, usant de figures de style sans but pragmatique);
> – on peut tout autant concevoir un texte en langue typologiquement latine et d'un niveau rhétorique très modéré (usage de phrases courtes, dont la structuration hypotactique ne dépasse pas deux ou trois niveaux, où l'ordre des mots est motivé par des soucis plus pragmatiques qu'ornementaux).

La lecture des textes montre néanmoins que, depuis le milieu du VII^e^ s. où apparaissent des passages protoromans, ces derniers sont écrits dans un souci pragmatique qui implique l'usage d'une rhétorique modérée. La corrélation semble cependant de moins en moins systématique, comme on l'observe nettement dans les chartes à partir du milieu du VIII^e^ s. Il y a là une piste qui conduit à affiner le modèle de Michel Banniard: cela ne sera pleinement possible qu'à l'aune du modèle pertinent pour la période qui commence au X^e^ s., et qui reste à construire.

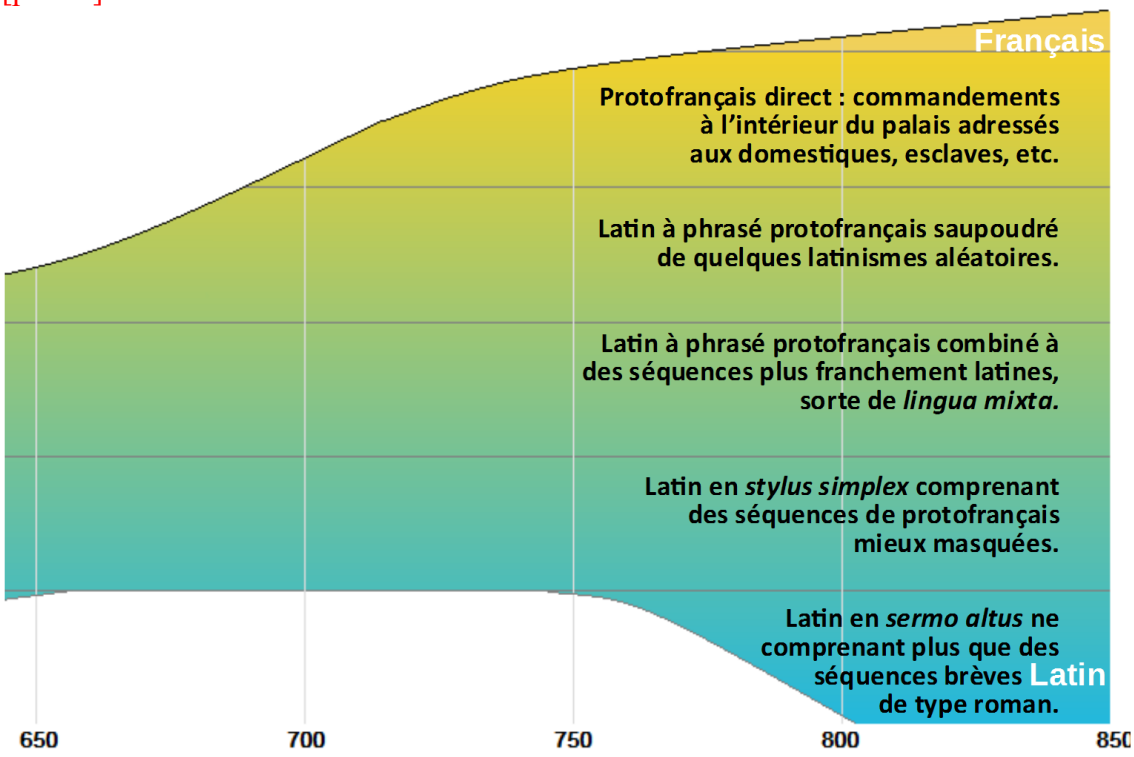
Il est en tout cas d'ores et déjà possible, pour le mettre en œuvre concrètement, d'articuler ce modèle carolingien à la périodisation proposée par le même M. Banniard[7], périodisation dont la pertinence est vérifiée dans notre thèse pour la période courant de 650 à 850. Ainsi, le modèle peut se représenter en diachronie. On voit qu'il est pleinement pertinent à partir de l'an 800 environ, époque où le diasystème latin achève une phase de dilatation rapide. L'évolution naturelle de la langue, concomitante à la fin du VIII^e^ s. à la restauration d'une langue traditionnelle inspirée de la prose patristique[8], a amené le

---

[7] BANNIARD M. (n. 1), 485–493.

[8] "Pépin le Bref, après 750, souhaita que la langue des diplômes qu'il émettait fût plus proche de la langue traditionnelle, avec un certain résultat. Ce fut le début d'une évolution qui ne fut accomplie que sous Louis le Pieux. Cette volonté fut encore plus nette chez Charlemagne, surtout après la conquête du royaume lombard en 774" (BOURGAIN, P. – HUBERT, M.-C.: *Le latin médiéval.* Turnhout 2005, 52), et plus encore, pourrait-on dire, après la venue d'Alcuin à la cour du roi en 782.

diasystème à se diffracter, jusqu'à se dédoubler, alors que la graphie latine, seule existante[9], maintient une apparence d'unité langagière.
[p. 458]



Cette modélisation diachronique n'est pas le premier essai du genre. Elle s'inspire de deux graphiques: l'un représentant l'évolution du latin écrit et du latin oral de 400 av. J.-C. à l'an mille[10], et l'autre représentant, pour chaque époque allant de 200 av. J.-C. à l'an 2000, l'élévation moyenne du niveau stylistique de la production en langue latine[11]. Ces deux modèles comportent des limites très nettes qu'il n'est pas possible de présenter ici[12].

Dans la modélisation proposée, l'on comprend mieux l'enchaînement de faits riches d'enseignements sociolinguistiques, notamment la prescription du concile de Tours en 813[13] et l'apparition en 842 d'un texte roman par sa structure comme par son système graphique: la version française des serments de Strasbourg. L'écart entre la latinité la plus travaillée et la langue la plus spontanée ne représentait plus seulement une simple

[9] Les premières tentatives d'écriture dans une nouvelle graphie adaptée à la prononciation naturelle apparaissent, pour la future zone de langue d'oïl, à la fin du VIII[e] s.: "en Gaule (HERMAN 1990b, 147), on peut remonter jusqu'à la Parodie de la loi salique, *et ipsa cuppa frangant la tota,* qui date d'environ 751-768 (AVALLE 1965a; 1965b), mais qui ne représente pas encore une francophonie comparable à celle des Serments de Strasbourg (842)." (INAICHEN, G.: L'apparition du roman dans des contextes latins. In FRANK, B. – HARTMANN, J. – SELIG, M. (éds): *Le passage à l'écrit des langues romanes.* Tübingen 1993, 84).

[10] PULGRAM, E.: Spoken and written Latin. *Language: Journal of the Linguistic Society of America* 26 (1950) 462.

[11] BERSCHIN, W.: *Biographie und Epochenstil im lateinischen Mittelalter*. *III: Karolingische Biographie, 750–920.* Stuttgart 1991, 148.

[12] Les critiques sont présentées dans la thèse: cf. n. 2, 406–409.

[13] *... et ut easdem omelias quisque aperte transferre studeat in rusticam Romanam linguam aut Thiotiscam, quo facilius cuncti possint intellegere quae dicuntur* (Concile de Tours, c. 17, *MGH Legum sectio* III, *Concilia* II, 1. 288).

variation stylistique, mais la traversée de plusieurs registres de langue. Sans en avoir encore bien conscience, ceux qui maîtrisaient tout le diasystème connaissaient en fait deux systèmes linguistiques imbriqués. Comment les ressources de ces deux systèmes s'articulent-elles au sein de ce *continuum*?

[p. 459] 3. Méthode et conclusions de la thèse

Le corpus analysé dans la thèse se compose de 21 chartes originales issues du fonds de l'abbaye de Saint-Denis, de 660 à 868, et de 3 *vitae* mérovingiennes réécrites à l'époque carolingienne:

	Vita s. Bathildis		*Vita s. Richarii*		*Vita s. Galli*		
Version	*I*^{a}	*II*^{a}	*I*^{a}	*II*^{a}	*I*^{a}	*II*^{a}	*III*^{a}
Dates	679–690/691	800–833	ca. 730	Alcuin, 800–804	ca. 680, 715/725, 771	Wettinus, ca. 820	Walafrid, ca. 833–834

Il s'agit de répartir un certain nombre d'éléments de la langue selon leur période d'existence dans le cours de la langue latine et de la langue romane. La typologie ci-dessous est inspirée des travaux de Michel Banniard, qui a publié en 1998 une première classification d'abord ternaire[14], avec des éléments "stables" (ici: "permanents"), "métastables" et "évanescents". Le même auteur a affiné ensuite sa typologie en introduisant la notion d'éléments innovants et en renommant les éléments "stables" en éléments "rémanents"[15]. Pour cette démonstration, l'étude très précise de quelques textes a conduit à étoffer cette typologie en introduisant deux autres types: un échelon supérieur d'innovation (éléments "très innovants") et une catégorie spécifique pour les éléments "transitoires", qui sont moins nombreux et pouvaient être négligés à première vue.

[14] BANNIARD, M.: Diasystèmes et diachronies langagières du latin parlé tardif au protofrançais, III^{e}–VIII^{e} siècle. In HERMAN, J. (dir.): *La transizione dal latino alle lingue romanze: atti della tavola rotonda di linguistica storica (Università Ca' Foscari di Venezia, 14–15 giugno 1996).* Tübingen 1998, 144.

[15] BANNIARD, M.: Action et réaction de la parole latinophone: démocratisation et unification, III^{e}–V^{e} siècles. *Antiquité tardive* 9 (2001) 127.

Voici donc les six types de structures:

“permanentes”: aussi fréquentes dans les textes latins traditionnels que dans les textes d’ancien français. Exemple: l’imparfait de l’indicatif comme dans *rogabat* > “revout”.
“très innovantes”: rares et/ou peu grammaticales en latin traditionnel, ou simplement tardives, elles ne seront pas complètement grammaticalisées avant l’ancien français. Exemple: *cum* [avec] remplacé par *apud* > ancien français “od”.
“innovantes”: rares et/ou peu grammaticales en latin traditionnel, ou simplement tardives, elles sont assez fréquentes pour être considérées comme grammaticalisées avant le VI^e^ s. et n’être pas particulièrement “marquées”. Exemple: *uir* remplacé par *homo*.
[p. 460] “transitoires”: rares et/ou peu grammaticales en latin traditionnel, ou simplement tardives, elles sont quasi-grammaticalisées, mais ont disparu avant l’ancien français. Exemple: *in corpore* ou *corporali modo* remplacent *corporaliter*, mais seront remplacés par *corporea mente* et *corporali mente* (en ancien français “corporement” et “corporeilment”).
“métastables”: disparues de l’oralité commune en Gaule du nord dans le courant du IX^e^ s., elles ne laissent que des traces résiduelles dans le très ancien français. Exemple: le plus-que-parfait “synthétique” de l’indicatif, comme dans *rogaverat* > “roveret”.
“évanescentes”: disparues de l’oralité commune en Gaule du nord dans le courant du VII^e^ s., elles ne laissent aucune trace en ancien français. Exemple: l’imparfait du subjonctif comme dans *rogaret,* remplacé par la forme *rogauisset* > “rovast”.

Chaque élément d’une phrase est étiqueté dans une des six catégories. La phrase concernée est ensuite elle-même placée dans un des cinq registres. Dans les exemples suivants, les éléments sont surlignés. Lorsque l’étiquetage concerne non pas les formes ou le lexique, mais l’ordre des mots, on a employé un soulignement:

Registre 5. – Protofrançais direct: notre corpus de chartes et de vies de saints n’en comporte aucun exemple. Voici un exemple tiré d’un autre type de texte, dépourvu de toute portée littéraire, et très terre à terre: la description de la basilique Saint-Denis en 799[16].	*(Basilica) habet de longo pedes CCLXV. De latus pedes CIII. De alto usque ad camerato habet pedes LXXV… In summo sunt intus illa ecclesia columnas inter totum XC. Excepto habet foras per illos porticos de illa ecclesia columnas capitales LVIII (…).*
Registre 4. – Latin à phrasé protofrançais saupoudré de quelques latinismes aléatoires: jugement royal rendu en 861[17].	*Dixerunt quod de presente tales testes idoneis colonis de predicta uilla Mintriaco abebant, per quem eis probare potebant quem, in tempore aui et genitori nostri bone memorie Hludouuici, ipsi et illorum antecessores*

[16] STOCLET, A. J.: La *Descriptio Basilicae Sancti Dyonisii*: premiers commentaires. *Journal des Savants* 1–2 (1980) 104.
[17] TESSIER, G. (éd.): *Recueil des actes de Charles II le Chauve, roi de France.* Paris 1955, n° 228.

	suprascripti serui ad imfreriorem seruicium de iam dicta uilla serper fuissent et plus per drictum et per legem quem coloni, sicut manifestum est, fecissent.
[p. 461] Registre 3. – Latin à phrasé protofrançais combiné à des séquences plus franchement latines, sorte de *lingua mixta*: *Vita Galli prima* (entre 680 et 771)[18].	*Inuenerunt ipsam pallulam cum ipsa cera inuolutam in fauillam cineris inconbustam, neque tetigit eos ignis [o]mnino, sed inlesa reperta sunt.*
Registre 2. – Latin en *stylus simplex* comprenant des séquences de protofrançais mieux masquées, sorte de *lingua mixta*: *Vita Bathildis secunda* (IX^e^ s.)[19].	*Erat enim benigna animo et moribus omnibus pudica, sobria, prudens et cauta, nulli machinans malum. Non leuis in eloquio, non praesumptuosa in uerbo, sed cuncta opera sua honestissimo moderabat ingenio.*
Registre 1. – Latin en *sermo altus* ne comprenant plus que des séquences brèves de type roman: *Vita Galli secunda* (IX^e^ s.)[20].	*Pallula una cum cera in cineribus inlesa reperta est, in quibus nec minimum quid ignis obsorbere ausus est, inaudito modo resistente cera incendio, quousque cremaretur in sancti Galli seruitio. Miraculum multis notum est, cum laus Christi in populis dilatata est.*

Bien sûr, tout n'est pas parfaitement étiquetable: deux exemples le montreront.

Tout d'abord, l'ampleur et le degré d'hypotaxe des phrases ne sont pas représentés ici par de la couleur. Ce sont des caractéristiques qui déterminent elles aussi les registres de langue, mais sans nécessaire corrélation avec la fréquence des éléments archaïques ou innovants. Autrement dit, une langue simple ne signifie pas une langue innovante: les deux exemples de protoroman le montrent. Pour l'un (registre 1), c'est une langue simple de faible ampleur et sans hypotaxe. Pour l'autre, c'est l'inverse: il s'agit d'une période, phrase ample typique de la rhétorique juridique.

Autre exemple: comment étiqueter ce qui n'existe pas morphologiquement? Ainsi, les déterminants en fonction d'article sont un trait protoroman: dans *per illos porticos* (registre 5), *illos* est étiquetable. Mais dans *in cineribus* (registre 1), l'absence de *illis,* et même de tout déterminant moins marqué (comme *eis*), pourrait être considéré comme un archaïsme et mériterait un étiquetage bleu foncé; or, il n'y en a pas.

Dernier exemple: la notation des désinences casuelles. Il est manifeste que les passages protoromans présentent beaucoup plus de désinences trahissant une prononciation naturelle, où des lettres se sont amuïes: ainsi dans *usque ad camerato* (au lieu de *cameratum*). On peut penser sans grand risque que la prononciation des passages protoromans n'était pas conservatrice. Mais l'usage d'une graphie latine a poussé les [p. 462] rédacteurs à utiliser encore assez souvent des désinences complètes dans les registres intermédiaires (surtout les registres 4 et 3), ce qui laisse un grand doute sur la prononciation et ne permet pas de présumer que, dans les registres 5 à 3, une terminaison en *-ibus* ait plus été prononcée que le *-s* pluriel en français actuel.

Malgré ces limites, l'effort d'étiquetage donne un aperçu déjà clair du fonctionnement des registres. L'intérêt d'étiqueter systématiquement un corpus de textes

[18] *Vita Galli confessoris triplex*. In *Monumenta Germaniae historica*. Éd. B. KRUSCH. Hanovre 1902 (Scriptorum rerum Merovingicarum 4: Passiones vitaeque sanctorum aevi Merovingici II), 255.
[19] *Vita sanctae Balthildis*. In *Monumenta Germaniae historica*. Éd. B. KRUSCH. Hanovre 1888 (Scriptorum rerum Merovingicarum 2: Fredegarii et aliorum chronica. Vitae sanctorum), 483.
[20] Cfr. n. 14, 279.

est de repérer les fluctuations de langage selon les genres textuels ou au sein d'un même texte. Ces fluctuations, peu arbitraires, montrent un usage très pragmatique des possibilités linguistiques de l'époque.

L'application de la grille sociolinguistique a ici toute sa pertinence à travers les concepts que voici:

* *diastratie*: les fluctuations peuvent apparaître selon l'appartenance sociale du rédacteur. Mais les concepts les plus opératoires ici sont les suivants:

* *diaphasie*: la fluctuation intervient surtout selon le niveau de langue de ceux à qui seront lus les textes;

* *diachronie*: la fluctuation devient rapidement bien plus large à partir de la fin du VIII[e] s. Le registre "protoroman direct" n'a pour l'instant pas été relevé avant cette époque.

Une fois ces relevés établis à la main, un important travail de synthèse s'est imposé pour proposer une description plus objective des ressources que privilégie ou évite chaque registre.

L'ensemble des phrases étiquetées "de tête" permet de dégager des tendances linguistiques parfois très nettes, formant un système articulable en 5 registres menant insensiblement d'un protofrançais latiniforme à du latin de facture traditionnelle. Voici un extrait du tableau de synthèse relevé, avec la répartition des éléments innovants dans les 4 registres trouvés dans les 21 chartes étudiées:

Registres	4	3	2	1
perfectum passif surcomposé (*amatus fuerat* remplaçant *amatus erat*)	+	+	+/–	+/–
nouveau subjonctif imparfait (*amavisset* > "amast", forme remplaçant *amaret*)	+	+	+/–	– –
passif analytique à l'*imperfectum* (*amatus erat/estabat* remplaçant *amabatur*)	– –	– –	– –	– –
passé analytique (passé dit composé: *amatum habet* remplaçant *amauit*)	–	– –	– –	– –
nouveau futur du présent en {infinitif+*habeo*} (*amare habet* remplaçant définitivement *amabit*)	– –	– –	– –	– –
nouveau futur du présent en {infinitif+*debeo*} (*amare debet* remplaçant provisoirement *amabit*)	++	++	–	– –
nouveau passé antérieur en {habui+participe passé}: *habui amatum* pour *amaueram*	– –	– –	– –	– –
[p. 463] nouveau plus-que-parfait en {habebam+participe passé}: *habui amatum* pour *amaueram*	– –	– –	– –	– –
adverbes en {adjectif+*mente*} (*corporali mente* remplaçant *corporaliter, corporali modo, in corpore*)	– –	– –	– –	– –
relatifs synthétiques au cas régime (*[ex]inde, unde* remplaçant *ex quo / ex qua / ex quibus*)	++	+	+/–	–
emploi de *id sunt* à la place de *ea sunt*	+	+	– –	– –

Un ordinateur est-il capable d'étiqueter toutes les phrases d'un corpus, de manière à flairer rapidement les variations linguistiques? Il se trouve que oui, et en créant une méthode autre, que seul un ordinateur peut mettre en œuvre.

4. Automatiser le repérage: application d'une méthode computationnelle

La méthode employée gagnerait à être appliquée et consolidée sur un corpus bien plus large. Cela permettrait bien sûr d'affiner le fonctionnement du diasystème. Surtout, le repérage de ces registres, articulé à leur cadre chronologique d'apparition, permettrait de proposer des hypothèses pour les textes non datés. Une fois leur situation d'énonciation bien posée, les différents registres de langue utilisés au sein d'un même texte, et la proportion dans laquelle ils y sont présents, pourraient ainsi servir d'indices chronologiques, voire géographiques.

L'ordinateur est-il capable d'étiqueter le niveau de langue d'un fragment de texte, sans qu'un humain procède à cet immense travail préalable qu'est l'étiquetage de chaque élément du fragment (degré d'innovation dans la morphologie verbale, dans le lexique, dans l'ordre des mots, etc.)?

4.1. Une problématique trop complexe pour les outils computationnels?

Pour appliquer nos analyses grâce à des outils informatiques, plusieurs démarches étaient *a priori* envisageables, mais ont rapidement été écartées ou abandonnées.

Nous aurions par exemple pu isoler les différentes formes caractéristiques d'un registre, en créant pour chacune d'entre elles des procédures spécifiques. Nous aurions ainsi compté les différentes formes présentes, en déduisant par additions et pondérations le registre de langue utilisé dans un fragment. À cette piste, séduisante par sa systématique, s'opposent cependant plusieurs obstacles.

Il est vrai qu'il serait aisé de surveiller la fréquence d'apparition de certains traits de vocabulaire, ou de diverses "expressions régulières". Par exemple, une tâche à laquelle un ordinateur excellerait serait de compter les occurrences d'éléments épidictiques comme *iamdictus, antefatus, memoratus, supra-/superscriptus, predictus* etc., ou d'adverbes comme *scilicet* et *uidelicet* (formes peu fréquentes dans le registre 4, et de plus en plus en fréquentes lorsque l'on se rapproche de la latinité traditionnelle). Déceler algorithmiquement certains glissements de sens requerrait cependant un travail [p. 464] beaucoup plus conséquent, avec des chances de succès probablement minces. Fréquente dans les registres 4 et 3, l'utilisation du plus-que-parfait du subjonctif avec le sens d'un imparfait du subjonctif (*ama[ui]ssem* signifiant *amarem*), est un phénomène auquel les procédures informatiques resteraient aveugles. Des confusions comme celles existant entre le subjonctif parfait (*amauerim, amaueris, amauerit*) et le futur antérieur (*amauero, amaueris, amauerit*) passeraient également inaperçues: la procédure informatique, tout comme certains locuteurs, ne ferait pas la différence entre les deux sens d'*amaueris* et *amauerit*. Ces erreurs potentielles ne sont pas anecdotiques. Or, dans de nombreux cas, rater un unique phénomène conduirait à un échec radical de la classification. En prenant pour objet, non le niveau de langue moyen d'un texte pris dans son intégralité, mais les variations de registres de langue au sein du texte, nous nous obligeons à étudier des unités relativement courtes, dans lesquelles le registre de langue peut n'être détectable que par un seul phénomène.

D'autres techniques auraient également pu nous bénéficier. Les méthodes dites de *classification non supervisée* permettent de regrouper automatiquement des passages qui se "ressemblent" le plus. La définition de la ressemblance est principalement laissée à l'ordinateur, d'où le nom de cette classe d'outils. Un algorithme de ce type réussirait-il à

percevoir les différences de niveau sans qu'on le guide? Si oui, confirmerait-il ou infirmerait-il nos propositions? Malheureusement, dans notre corpus, les différences de registre ne sont pas les seules discernables. Elles sont mêlées à des différences majeures dans les thématiques, dans les lieux, etc. Les classifications obtenues par des algorithmes comme les k-médoïdes[21] se sont ainsi révélées très faiblement pertinentes dans notre cas.

4.2. Reproduire notre regard: l'apprentissage par la machine

Face aux échecs de ces différentes méthodes, le choix de l'apprentissage automatique s'est finalement imposé. Si les mathématiques opérant dans ces méthodes sont complexes[22], l'idée sous-jacente à notre procédure est relativement simple. Nous étiquetons tout d'abord tous les passages de nos textes selon le(s) registre(s) de langue que nous y décelons lors de leur lecture cursive. Puis nous demandons à l'algorithme de trouver un calcul qui lui permette d'étiqueter lui-même les textes exactement de la même manière que nous. Il fournit alors un nouvel étiquetage, dont on peut comparer la proximité à l'égard du nôtre. Si les deux étiquetages sont suffisamment proches, on considère que la machine a bien appris à reproduire notre résultat par sa propre démarche. On peut alors lui fournir de nouveaux textes, que nous n'aurions pas annotés nous-mêmes au préalable.

[p. 465] De manière classique, on pourrait considérer les textes qu'on lui fournit comme un "sac de mots"[23]. Dans ce type d'approche, une phrase comme *incipit uita beatae Bathildis reginae* est transformée pour l'ordinateur en cinq objets séparés: *incipit* / *uita* / *beatae Bathildis* / *reginae*. L'ordre dans lequel ces mots sont présentés n'est pas une information que l'ordinateur retient dans ce cas. Le texte est considéré comme une liste de mots, apparaissant chacun un certain nombre de fois. Les passages à étudier étant relativement courts, il est toutefois essentiel que l'apprentissage soit fait en extrayant un maximum d'information des textes fournis. Cette approche canonique nous a donc paru insuffisante.

Pour préserver une partie de l'information concernant l'enchaînement des différentes formes présentes dans un texte, nous avons considéré les textes comme un ensemble de "n-grammes" de caractères. L'ordinateur ne stocke plus seulement une liste de mots, mais, pour notre exemple, dans le cas de 5-grammes de caractères, une liste des enchaînements de 5 caractères: "incip", "ncipi", "cipit", "ipit_", "pit_u", "it_ui", "t_uit", "_uita", "uita_", "ita_b", etc. Cette démarche présente également d'autres avantages. Elle permet en particulier à l'intelligence artificielle de faire la distinction entre les parties importantes du mot, et d'autres qui le seraient moins. Par exemple, dans le cas des noms propres: le lieu ou la personne évoqués importent peu pour définir le registre de langue. "*Bathild-*" n'est donc pas une information pertinente. Que la désinence soit "*-is*" n'est pas une information particulièrement décisive, mais elle est tout de même un élément

---

[21] KAUFMAN, L. – ROUSSEEUW, P. J.: Clustering by means of medoids. In DODGE, Y. (ed.): *Statistical data analysis based on L1 norm and related methods.* North-Holland 1987, 405–416.

[22] Après avoir utilisé des machines à vecteur du support (*support vector machine*, ou SVM) avec un certain succès, nous nous tournons désormais vers des modèles de régression logistique, dont les performances semblent encore supérieures. Sur les SVM, voir BOSER, B. E. – GUYON, I. M. – VAPNIK, V. N.: A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on computational learning theory*. New York 1992, 144–152.

[23] ZHANG, Y. – JIN, R. – ZHOU, Z.-H.: Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics* 1 (2010) 43–52.

constitutif de l'évaluation de langue utilisée dans le texte, que le découpage en n-grammes de caractères permet d'isoler.

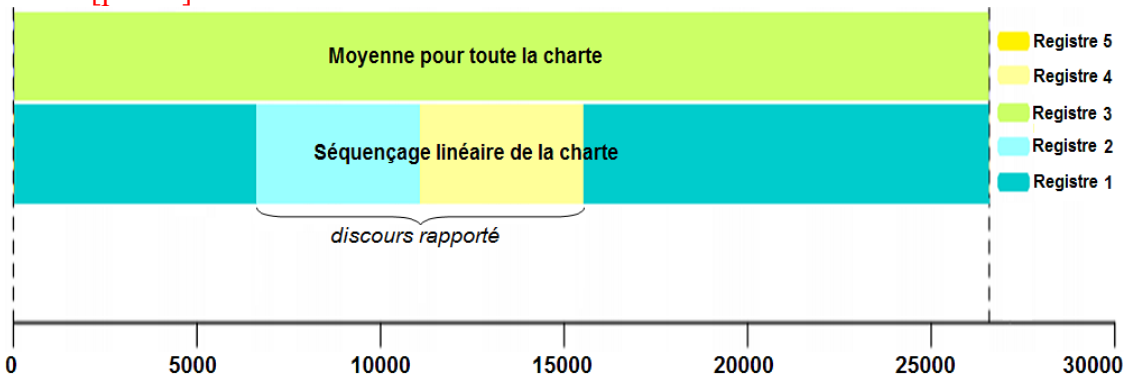
Nous avons choisi d'entraîner notre intelligence artificielle à reproduire nos annotations, en lui faisant percevoir le texte comme une succession de tous les fragments de 60 mots possibles, analysés par paquets de 5-grammes de caractères. Ces valeurs apparemment arbitraires sont en réalité le fruit de tâtonnements successifs. Elles correspondent, en l'état de notre recherche, à la configuration ayant donné les meilleures performances.

4.3. Des résultats encourageants

Les premiers résultats obtenus grâce à cette méthode sont très encourageants. Au-delà de résultats très satisfaisants sur les textes aux registres de langue homogènes, notre méthode a notamment prouvé sa capacité à éviter certains pièges, en repérant des structures innovantes au sein de textes en latin traditionnel.

Le texte étudié ici est la charte de 861, formant près de 3400 caractères, découpés par unités de 60, où sont encore découpées des séquences de 5 caractères, selon la méthode exposée plus haut.

[p. 466]



On obtient plus de 26000 caractères: c'est l'axe des abscisses.

On voit aisément que le début et la fin de la charte sont d'une latinité conservatrice: l'ordinateur y décèle le registre de langue le plus conservateur (registre 1), là où, de tête, nous avons étiqueté les registres 1 et 2.

Au milieu de la charte, là où le formulaire est moins suivi car cette partie met en œuvre un discours rapporté, l'ordinateur trouve du registre 2, puis du registre 4, là où nous avons étiqueté un registre 3 puis un registre 4.

L'ordinateur a donc perçu les variations de registre dans les grandes lignes, de manière encore approximative pour les registres latinisants, et de manière quasi-exacte pour le registre protoroman. Le passage protoroman en question correspond à l'extrait présenté plus haut pour illustrer le registre 4 (*dixerunt quod de presente…*).

5. Conclusion

L'entraînement de notre intelligence artificielle n'en est qu'à ses débuts: celle-ci sera encore plus fiable avec de nouveaux textes d'entraînement, pré-analysés par l'esprit humain, jusqu'à ce que de nouveaux ajouts deviennent négligeables. Il n'en reste pas moins que ces premiers résultats ouvrent des perspectives enthousiasmantes, car l'analyse

computationnelle des textes anciens reposait jusqu'ici sur un étiquetage préalable par l'humain, souvent avec l'insertion manuelle de balises XML pour chaque mot et/ou groupe de mots, ce qui correspond à un travail préparatoire aussi volumineux que le corpus choisi. Inversement, la méthode de *clustering* et d'apprentissage automatique présentée ici affranchit à moyen terme le linguiste de ce lourd préalable.

Florian Cafiero
LIED (UMR 8236) – Université Paris Diderot / Université Sorbonne Paris-Cité/CNRS
10 rue Alice Domon et Léonie Duquet – 75013 Paris
France
cafiero@phare.normalesup.org

Rémy Verdo
Ecole nationale des chartes, PSL Research University, Centre Jean-Mabillon – Archives municipales de Toulouse
remy.verdo@orange.fr