PICARTEXT : Une ressource informatisée pour la langue picarde

Jean-Michel Eloy, Fanny Martin, Christophe Rey

PICARTEXT : Une ressource informatisée pour la langue picarde

Jean-Michel Eloy[1], Fanny Martin[1], Christophe Rey[1]

(1) LESCLAP (CERCLL-EA 4283), Université de Picardie Jules Verne, Amiens
jean-michel.eloy@u-picardie.fr, fanny.martin@u-picardie.fr, christophe.rey@u-picardie.fr

Résumé.

Picartext est une base de données textuelles, construite depuis près de 10 ans à l'Université de Picardie à Amiens. Elle présente des caractéristiques de premier intérêt pour la recherche sur les traitements automatiques. La langue picarde, d'une vitalité non négligeable, dispose d'une littérature assez abondante et de très nombreux dictionnaires et glossaires. Mais elle ne possède pas de standard, ni linguistique, ni graphique. La langue est donc très variante. La base de données, de nature littéraire, d'environ 5 millions d'occurrences, est accessible en ligne au moyen d'un outil d'interrogation paramétrable : non seulement il permet la restriction du corpus de travail (lieux, dates, genres), mais il permet une recherche tenant compte d'équivalences phonétiques et d'équivalences dialectales. Il est ouvert à des évolutions en termes de balisage, en particulier dans le cadre d'un projet ANR portant sur trois langues régionales simultanément (picard, alsacien, occitan).

Abstract.

PICARTEXT : a computerized resource for picard

Picartext is a textual database, built up since about 10 years in Picardy University in Amiens. Some of its characteristics make it very interesting for research on natural languages processing. Picard language, of a not insignificant vitality, has a rather plentiful literature, and very numerous dictionaries and glossaries. But it does not possess standard, either linguistics, or graphic. The language is thus very variant. The database, of literary nature, counts about 5 million token, is reachable on-line, with a customizable tool of interrogation : not only it allows the limitation of the working corpus (places, dates, genres), but he allows a search taking into account phonetic equivalences and dialectal equivalences. It is opened to evolutions in terms of tagging, in particular within the framework of an ANR project concerning three regional languages simultaneously (picard, alsatian, occitan).

Mots-clés :

picard, non standardisation, variation dialectale, variation graphique, numérisation, balisage, équivalences

Keywords:

picard language, non standardisation, dialectal variation, graphical variation, digitisation, tagging, equivalences

1 La langue picarde

1.1 Données générales sur le picard, langue de France

La langue picarde se trouve sur un territoire vaste, mais divisé puisqu'elle s'étend sur deux régions françaises et une région de Belgique. En dépit de cette situation géographique, son originalité et son unité sont assez fortes. La carte reproduite ci-dessous permet de se figurer son domaine linguistique :

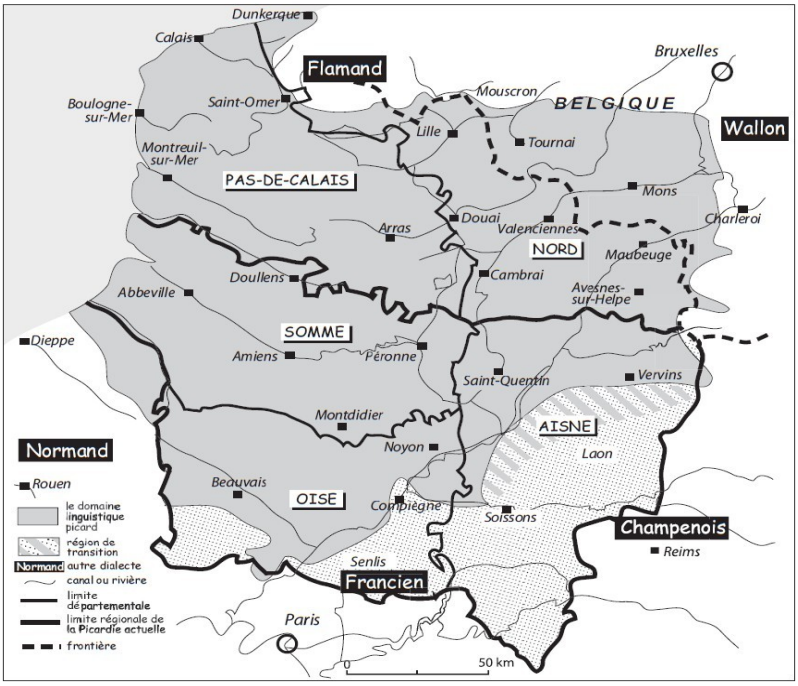

FIGURE 1: L'aire linguistique picarde, carte établie par René Debrie selon les informations de Raymond Dubois, réalisée par Joëlle Désiré pour l'Atlas de Picardie (Amiens, Université de Picardie Jules Verne)

La région connaît un assez fort investissement identitaire, mais il existe une coupure entre *picard* et *patois*, *chtimi*, *rouchi*[1]. Le statut de la langue, encore discuté aujourd'hui, est marqué par une reconnaissance incomplète par les pouvoirs publics[2]. En effet, alors que cette reconnaissance est moyenne en région Picardie – dans la mesure où le picard peut s'appuyer sur l'existence de discours de promotion, notamment financés par des fonds publics – elle est en revanche très faible en région Nord-Pas-de-Calais, comme au niveau national, et ce malgré le rapport Cerquiglini de 1999.

La langue picarde appartient au groupe gallo-roman, dans une relation de proximité dynamique « langues collatérales » (Eloy 2004) avec le français. C'est une « variété basse de diglossie » assez typique – et qui fait l'objet actuellement d'efforts de « retroussement de la diglossie » (Lafont 1984).

1.2 Un développement sans standardisation

L'historicité de la littérature en langue picarde est considérable. Les pratiques littéraires dans cette variété s'étendent en effet depuis le Moyen-Âge (dans le cadre d'une langue d'oïl tolérante), se renforcent au XVII[e] siècle, et connaissent ensuite une croissance constante au XIX[e] siècle. Les publications actuelles se comptent quant à elles par dizaines chaque année.
En dépit de cette histoire littéraire déjà longue et riche, le picard est caractérisé par un développement sans standardisation, qui semblerait se faire au bénéfice de « pôles de pratiques » (Forlot & Martin, 2014 ; Martin, 2015; Martin & Forlot, à paraître).

---

1 Bien qu'il s'agisse de la même langue, il existe néanmoins des variétés. Ces appellations correspondent à des nominations différentes sur le continuum géographique.

2 Mais également par les habitants du domaine linguistique, locuteurs ou non locuteurs du picard.

Considérons par exemple ce qui se passe pour la production lexicographique dans cette langue. Le picard jouit d'une richesse lexicographique sans égale par rapport aux autres langues régionales de France puisque l'on peut recenser plusieurs centaines de titres proposés depuis le XVIIIe siècle jusqu'à nos jours. Existant en l'absence d'une standardisation, cette richesse s'explique à la fois par la tradition dialectologique, une forme d'aménagement du statut (de nombreux lexiques et glossaires se trouvent ainsi insérés en fin de volumes de récits) et une « grammatisation militante » : dictionnaires français-picard, dictionnaires en ligne.
Bien sûr, cette situation de non standardisation est liée au statut de la langue qui ne bénéficie d'aucun enseignement (Forlot & Martin, 2015), d'aucun usage officiel, ni de politique linguistique véritable. Et même actuellement, il n'y a aucune demande de la part des acteurs du monde picardisant pour faire évoluer cette configuration.

Intéressons-nous ensuite à la question de l'orthographe en picard, dimension qui confirme elle aussi l'absence d'une standardisation, malgré de nombreux débats antérieurs.
L'orthographe picarde voit la coexistence de quelques systèmes et d'une marge anarchique. Citons en exemple la séquence « c'était » que l'on peut ainsi retrouver sous les formes *ch'étoué, ch'étouait, ch'étouis, ch'étouo, ch'étwo, ch'étouot, ch'étwot,* et même *ché toué*, etc. Cette multiplicité de formes traduit certes des phénomènes de variation dialectale à l'intérieur de l'espace linguistique picard, mais elle atteste également une tendance forte à la variabilité graphique et à des problèmes de segmentation possiblement imputables à l'absence de standard, phénomènes particulièrement manifestes dans la base PICARTEXT.
Un second exemple illustrant la grande richesse orthographique du picard peut être donné à travers l'évocation du *Dictionnaire général français-picard* de Jean-Marie Braillon, ouvrage de 2001 dans lequel sont en effet recensées pas moins de 18 graphies différentes pour le mot « aiguille ».

L'un des points les plus saillants de la variabilité orthographique du picard (Dawson, 2002) concerne sans doute l'utilisation des apostrophes qui permet de dégager des séquences aussi distinctes que les suivantes: *chol vaque, chov vaque, chov'vaque, cho'v vaque/ch'timi, chtimi/pèmes tère, pèm'terre*, etc.

En pratique, on retiendra qu'il existe quatre grands types de graphies pour la langue picarde - dont trois constituent des substandards -, à savoir l'orthographe proposée par Vasseur (Vasseur, 1968), celle proposée par Feller et Carton (Carton, 2001), celle livrée par Braillon (Braillon, 1991), et enfin les cacographies. Soulignons toutefois que de récents succès de librairie, dont une traduction des albums d'Astérix tirée à 130 000 exemplaires, ont vu la mise en évidence de substandards.
En bref, malgré l'historicité, et la nette unité, la caractéristique centrale en picard est la variation maximale : une variation en liberté, qui est aussi une forme de contrainte dans son expansion et se traduit par une « double insécurité » (Martin, 2015) chez les locuteurs et les néo-locuteurs.

2 PICARTEXT

Nous livrons ci-dessous un aperçu synthétique de la base de données PICARTEXT conçue au LESCLAP, en esquissant successivement une description de la nature du corpus textuel mis en place et en apportant ensuite des éléments d'information concernant les modalités techniques d'informatisation et d'exploitation de celle-ci.

2.1 Le corpus Picartext

La base de données[3] a été conçue en fonction de problématiques linguistiques, et non pas purement pour le Traitement Automatique des Langues. C'est un travail qui se situe dans la continuité du Centre d'Etudes Picardes (CEP) - hébergé désormais par le laboratoire *Linguistique Et Sociolinguistique : Contacts, Lexique, Appropriations, Politiques (LESCLAP)*. L'équipe picarde travaille par ailleurs sur les problématiques de langues minorées, de langues proches, et de politiques linguistiques.

Nous avions dans l'équipe une connaissance d'utilisateurs de Frantext, et une expérience de travaux de rétroconversion de dictionnaires anciens. Des études de faisabilité ont été réalisées par des étudiants de Lille inscrits dans un Master de lexicologie et lexicographie. En tant que linguistes, nous avions une bonne connaissance de la langue et de la littérature

[3]La base PICARTEXT est constituée de textes écrits partiellement ou totalement en picard, issus de l'ensemble du domaine linguistique picard, et composés depuis le XVIIIe siècle jusqu'à nos jours. Son objectif est d'offrir à la communauté des chercheurs, ainsi qu'à un public averti, une ressource linguistique à partir de laquelle il sera possible d'envisager toutes sortes d'exploitations :
- exploration de la langue (lexique, morphosyntaxe, phraséologie...)
- étude des évolutions diachroniques
- étude de la variation et de la cohésion dialectales et du processus de koinèisation

picardes. Deux informaticiens-linguistes ont ensuite, grâce à des contrats postdoctoraux, réalisé la base. (Eloy-Rey-Dawson, 2011)

Le contenu de la base est un corpus de littérature d'environ cinq millions de mots. Il est difficile de dire ce que représente la totalité des publications en picard : 80 millions d'occurrences ? 150 millions ? Plus ? Notons que la question est insoluble aussi, à une autre échelle, en français.
Le problème de la sélection des textes, malgré un effort de rationalisation, a été réglé empiriquement (sur la base de « noms bien connus »). Il semble d'ailleurs que le grand modèle qu'est Frantext[4] n'ait pas pu procéder autrement. La difficulté augmente avec le développement de la littérature, donc elle est plus grande pour le XX^{e} siècle que pour le XVIII^{e} siècle.
La diversité est très grande, saisie en termes de lieux, de dates et de genres. Les genres, par exemple, jouent diversement sur la qualité de la langue : imitation du parler, imitation du français, lyrisme, vulgarité, etc.
La notion d'œuvre (ou même de texte) pose quelques problèmes. Le premier est celui de l'alternance, car il existe des textes alternants, ou bilingues, ou simplement des préfaces en français. Pour dégager un corpus seulement en picard, nous avons donc été dans l'obligation de sous-délimiter au sein des œuvres. Les scholies de théâtre, par exemple, sont parfois en français, parfois en picard, ce qui rend les choix délicats concernant ce type de « paratexte ». Plus embarrassante encore est la langue intermédiaire de certains textes, constituant une sorte de continuum souvent nommé « patois » (et non langue). Bref, la notion d'œuvre en picard implique des interventions sélectives, en partie arbitraires car dans cette culture les langues vivent ensemble et nous voulons les séparer.

2.2 Le processus de numérisation

La constitution de la base de données PICARTEXT s'est appuyée sur la collecte de quelques textes déjà numérisés, mais a surtout consisté à intégrer des textes majoritairement édités sur papier. Nous avons pour cela mis au point un itinéraire du texte édité, qui passe par plusieurs étapes. Les documents sélectionnés ont d'abord été photocopiés avant d'être traités par un logiciel de reconnaissance automatique de caractères. La qualité du rendu a tout de même nécessité une importante phase de double relecture des textes informatisés avant leur livraison sous un format .TXT considéré comme exploitable pour la suite des traitements informatiques et notamment la phase d'application d'un balisage logique au format XML. Le schéma reproduit ci-dessous illustre le parcours des textes intégrés dans PICARTEXT :

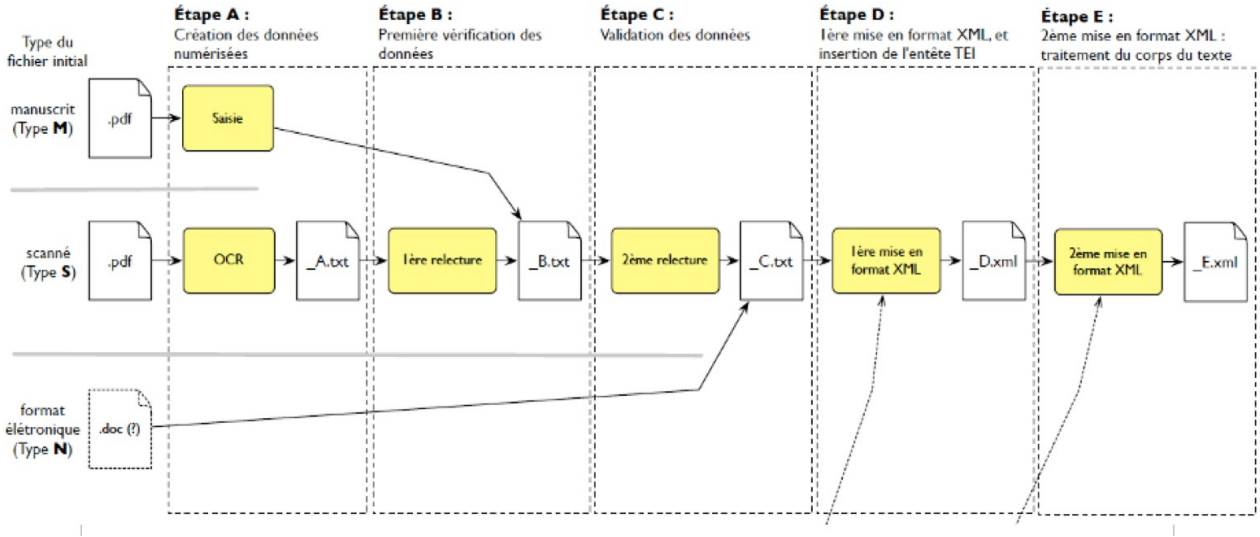

FIGURE 2 : Parcours des textes dans PICARTEXT

2.3 Description du module d'interrogation

La base PICARTEXT, désormais accessible en ligne par le biais d'un portail internet dédié[5], bénéficie d'un module d'interrogation répondant à des finalités du grand public et des chercheurs.

4 http://www.frantext.fr/
5 https://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/

Sans être obligatoire, l'interrogation de la base PICARTEXT permet de déterminer ou de restreindre un corpus de travail, selon plusieurs paramètres : lieux, dates, genres. Puis vient le choix d'une méthode d'interrogation.
Outre la recherche par « chaînes de caractères » et celle par « expressions rationnelles », on peut procéder aussi par deux méthodes plus élaborées, appuyée sur Dawson (2006), utilisant l'approche théorique de Mc Carthy & Prince (1995) :
- par « correspondance phonétique » : le mot est d'abord converti en sa représentation phonétique à l'aide d'un phonétiseur[6]. C'est cette représentation phonétique qui est recherchée, ce qui permet de ne pas tenir compte de l'orthographe des auteurs.
- par « correspondance dialectale » : le mot est converti en une forme abstraite (lemme dialectal) qui neutralise la variation dialectale du picard. Ceci permet de retrouver le mot sous diverses formes dialectales.
Le site donne des exemples de ces différentes modalités de requête.

La publication des résultats se fait actuellement par le biais de concordances récupérables sous la forme de données tabulaires ou au simple format texte.

2.4 Une base configurée pour être développée : la voie incontournable du balisage XML

Il est toujours à craindre que l'investissement humain et financier ne permette pas immédiatement de pousser les travaux aussi loin qu'on le voudrait. Notre stratégie a donc consisté à assurer d'abord la structuration du matériau linguistique - qui reste toujours enrichissable et continue d'ailleurs à s'enrichir - et les outils d'interrogation à disposition du public spécialiste ou non. Les standards TEI[7] et XML[8] ont ainsi été retenus, notamment dans une perspective de pérennisation de la ressource.
Le projet PICARTEXT s'est ainsi intégré à la galaxie des très nombreuses ressources pouvant bénéficier des travaux de balisage dans le langage XML. L'étape à laquelle nous arrivons est précisément celle de la mise en place d'un protocole de balisage logique XML conforme à la TEI qui puisse tenir compte de la grande diversité des types de documents (poésies, recueils de prose, pièces de théâtre, chansons, dictionnaires, etc.) constituant la base.

L'extension qualitative de la base nous amènera à y intégrer davantage de textes anciens. Nous avons aussi commencé à nous interroger sur la possibilité d'y ajouter des transcriptions d'oral. Ces différentes perspectives d'évolution nous amèneront nécessairement à parfaire le module d'interrogation lui-même et à le rendre capable de gérer au mieux l'accroissement des possibilités de recherche.

3 Les paris du projet RESTAURE

Depuis peu, le LESCLAP est l'un des partenaires scientifiques du projet de recherche ANR RESTAURE. Le LESCLAP avait pris l'initiative de se confronter à d'autres langues de situations similaires lors d'une journée d'étude qui préfigurait le projet RESTAURE, concrétisé grâce à des collègues travaillant sur l'alsacien et d'autres sur l'occitan.

Visant à « fournir des ressources informatiques et des outils de traitement automatique pour trois langues régionales de France : alsacien, occitan et picard », le projet RESTAURE repose sur le développement « de nouveaux modèles adaptés aux langues disposant de peu de ressources et peu standardisées ». Les procédures d'enrichissement des ressources, comme les traitements, devront prendre en compte la non standardisation, la variation linguistique et graphique, voire les faibles moyens financiers. *In fine*, l'objectif est de disposer de ressources textuelles, de corpus annotés et à partir de là, de lexiques morphologiques étendus, tout d'abord dans les trois langues envisagées, avec extension des méthodes à d'autres langues « peu dotées ». L'utilité sociale concerne donc les communautés linguistiques autant que les milieux de la recherche.

Le LESCLAP, fort de l'expérience de PICARTEXT, voit dans ce projet une opportunité de prolonger les acquis de cette ressource. Un des objectifs de ce programme, en 2008, était d'intéresser des informaticiens aux traitements de cette langue, et nous notons avec satisfaction, ainsi qu'en témoigne le nombre de visites du portail internet l'hébergeant, que la base jouit d'un intérêt véritable.

---

[6] Le phonétiseur utilisé dans le module expérimental de recherche dans le corpus Picartext est issu du système TTS-French développé par David Haubensack, sur la base des travaux de Thierry Dutoit, dans le cadre du projet MBROLA de la Faculté Polytechnique de Mons (Belgique). Références :
- Dutoit, T., V. Pagel, N. Pierret, F. Bataille, O. van der Vreken, 1996. "The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes" Proc. ICSLP'96, Philadelphia, vol. 3, pp. 1393-1396.
- Dutoit, Thierry. 1997. *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht Hardbound.
[7] TEI Consortium, <http://www.tei-c.org/Guidelines/P5/>.
[8] World Wide Web Consortium, <http://www.w3.org/TR/xml/>

Les attentes et les objectifs du projet RESTAURE, en ce qui concerne plus particulièrement la langue picarde, sont nombreux. Ce projet est par exemple une occasion d'étendre le nombre de textes en langue picarde mis à disposition sous forme numérique. Grâce à la mise en place d'un processus de numérisation améliorant fortement les réalisations de PICARTEXT, nous serons en mesure de détenir un dispositif méthodique et pérenne de constitution de nouvelles données.

Une des particularités très fortes de la langue picarde réside dans l'extrême variation graphique à laquelle elle est sujette. Le projet RESTAURE visant à mettre au point des outils de maîtrise de cette variation par la proposition de règles d'équivalence, nous nous interrogeons sur la possibilité de trouver une ou plusieurs solutions véritablement satisfaisantes. Ces règles pourront-elles d'ailleurs être éventuellement transversales aux trois langues du projet ou resteront-elles propres à chaque langue ? Voilà l'un des enjeux forts de RESTAURE.

Une caractéristique commune de l'alsacien, de l'occitan et du picard, est qu'il n'existe pour aucune de ces langues de dictionnaire faisant office de « standard ». Les outils automatiques de désambiguïsation grammaticale et lexicale envisagés dans le cadre du projet devront donc s'appuyer sur des procédures d'étiquetage morphosyntaxique particulièrement robustes. Il s'agit là d'une perspective particulièrement stimulante qui pourra notamment s'appuyer sur l'utilisation des dictionnaires de langues proches, par exemple allemand standard et français standard. On notera donc que le bénéfice ira aux méthodes, donc à toutes les langues, y compris celles déjà « bien dotées ».

Conclusion

Arrivé au terme de son financement, le projet PICARTEXT s'impose comme une ressource linguistique de tout premier ordre pour le picard. Mettant en évidence une langue qui se construit en liberté dans la variation et en absence de standardisation, ce projet doit encore être valorisé auprès du grand public pour atteindre les objectifs de sa création.
En ce qui concerne la communauté scientifique, PICARTEXT comble déjà toutes nos attentes puisqu'il offre non seulement des perspectives de recherche conséquentes, mais fait aussi entrer le picard dans le cercle restreint des langues régionales de France disposant, grâce à l'informatique et au Traitement Automatique des Langues, d'une visibilité accrue. L'intégration récente de l'équipe LESCLAP au projet RESTAURE constitue, selon nous, une illustration de l'intérêt de cette langue pour les recherches sur les autres variétés régionales et les langues plus richement dotées. Bénéficiant des initiatives audacieuses de ce projet, la base PICARTEXT devrait, sans nul doute, connaître dans les années futures des avancées considérables.

Remerciements Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet RESTAURE - convention ANR-14-CE24-0003-01) et du Conseil régional de Picardie.

Références

BRAILLON J.-M. (2001). *Dictionnaire général français-picard*, Tome I, éditions F.I.P.Q.

BRAILLON J.-M. (1991). *La graphie FIPQ du picard*, Lemé.

CARTON F. (2001). « Orthographe picarde Feller-Carton », in Linguistique Picarde, décembre.

CERQUIGLINI B. (1999). *Les Langues de la France, Rapport au Ministre de l'Éducation nationale, de la Recherche et de la Technologie*.

DAWSON A., ELOY J.-M., REY C. (2011 – non publié). Vue perspective sur le français à partir d'une base de données textuelles en domaine d'oïl, *Colloque annuel de l'Association for French Language Studies*, 8-10 septembre 2011, Nancy.

DAWSON A. (2006). *Variation phonologique et cohésion dialectale en picard. Vers une Théorie des Correspondances Dialectales*, Thèse de doctorat sous la direction de Marc PLÉNAT. 340 pages.

DAWSON A. (2002). « Le picard, langue polynomique, langue polygraphique ? », in D. Caubet, S. Chaker, J. Sibille (éd.*), La codification des langues de France*, L'Harmattan, Paris.

DUTOIT T., PAGEL V., PIERRET N., BATAILLE F., VAN DER VREKEN O. (1996). « The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes », Proc. ICSLP'96, Philadelphia, vol. 3, 1393-1396.

DUTOIT T. (1997). *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht Hardbound .

ELOY J.-M., REY C. (2012 – non publié). Une base de données lexicale en picard : la base PICARTEXT, *Journée d'étude du LESCLAP : "Bases de données informatisées dans les "petites langues""*, Amiens, 07 décembre 2012.

ELOY J.-M. (2004). *Des langues collatérales*, Paris, L'Harmattan.

FORLOT G., MARTIN F. (2014). « Entre invisibilité et (auto)occultation. Les paradoxes des pratiques langagières minoritaires en Picardie », in K. Djordjevic (éd.), Les minorités invisibles : diversité et complexité (ethno)sociolinguistiques, Éditions Lambert-Lucas, Limoges, pp. 77-87.

FORLOT G., MARTIN F. (2015). « Le picard à l'épreuve du terrain scolaire aujourd'hui », Communication dans le cadre du Congrès international de sociolinguistique, Grenoble, 2015.

LAFONT R. (1984). Pour retrousser la diglossie. *Lengas*, 15, 5-36.

MARTIN F., FORLOT G. (à paraître). « Hétérogénéité linguistique et poids des idéologies sur les pratiques linguistiques en Picardie », in A. Boudreau et L. Arrighi, *La construction discursive du locuteur francophone en milieu minoritaire. Problématiques, méthodes et enjeux*, Presses de l'Université Laval, Ste Foy (Québec).

MARTIN F. (2015). *Espaces et lieux de la langue en Picardie au XXIème siècle. Approche complexe de la structuration des répertoires linguistiques en situations ordinaires. Enquête en Picardie*, Thèse de doctorat, Université de Picardie Jules Verne, Amiens.

MCCARTHY J., PRINCE A. (1995). Faithfullness and Reduplicative Identity, *in* J. Beckman, L. Walsh Dickey, S. Uurbanczyk (éd.), *Papers in Optimality Theory*, U. of Massachusetts Occasional Papers in Linguistics 18, Amherst, Mass. : Graduate Linguistic Student Association, 249-384.

VASSEUR G. (1968). « L'orthographe picarde, principes généraux et règles pratiques établis par les Picardisants du Ponthieu et du Vimeu », *Linguistique Picarde*, décembre 1968 (graphie analogique).