

PICARTEXT : une expérience de base de données en langue régionale

Introduction

L'idée de l'élaboration de la base de données PICARTEXT revient au professeur Jean-Michel Éloy, chercheur en sociolinguistique et dont les nombreux travaux ont permis ces dernières années de mieux faire connaître encore la langue picarde. Dotée de longues et riches traditions littéraires et lexicographiques, cette variété régionale méritait en effet toute l'ambition d'un tel projet qui nous permet aujourd'hui de contribuer à ce volume. En ce sens, bien qu'absent de l'écriture de cette contribution, notre collègue ne peut qu'être étroitement associé à celle-ci.

La base PICARTEXT se trouve présentée ici non plus sous le seul angle scientifique¹, mais sous celui de la vulgarisation : des principes méthodologiques qui ont permis sa construction, en passant par la dimension proprement technique qui fait d'elle une base de données informatisée selon les plus rigoureux principes des grandes bases textuelles du monde de la recherche, et en allant jusqu'à l'évocation de ses retombées scientifiques actuelles.

1 — Cf. Éloy, J.-M., Martin, F., Rey, C. (2015) et Dawson A., Éloy J.-M., Rey C. (2011).

1. Une base de données littéraire à vocation représentative

S'appuyant sur la grande quantité d'œuvres littéraires disponibles dans la tradition linguistique picarde, et donc à l'instar de qui existe pour la base FRANTEXT, PICARTEXT est une base de données à forte coloration littéraire. Conformément à un souci de représentation la plus large possible des différents genres, la ressource associe à des romans en prose des textes en vers, mais aussi du théâtre, des chansons, ou encore des contes.

À ces textes, viennent également s'ajouter – en plus petit nombre – d'autres œuvres telles que des dictionnaires et des textes journalistiques. Le corpus étant amené à s'étendre dans les années futures, c'est notamment la part de ces documents qui devrait être accrue.

Au-delà du premier critère de représentativité typologique, nous avons également veillé à ce que la base constituée puisse rendre au mieux compte de l'ensemble de la variation connue par la langue picarde sur son vaste domaine linguistique et pas seulement sur son territoire administratif². Les textes constituant la ressource s'échelonnent donc sur les trois départements de la Somme, de l'Aisne et de l'Oise et vont jusqu'au Hainaut belge. L'accès plus ou moins facile aux documents de chaque aire nous a amenés et nous amène encore à procéder à des réajustements au cours de l'évolution de la base afin d'éviter des sur-représentations de variétés au détriment des autres.

Le dernier critère ayant présidé à l'élaboration de la base, non dissocié des deux précédents mais concomitant, est celui de l'inscription des textes de PICARTEXT dans une diachronie représentative de l'évolution de la langue picarde au fil des siècles. Dans cette perspective, nous avons donc collecté des textes s'échelonnant depuis le siècle des Lumières jusqu'à nos jours. Ce critère est particulièrement important puisqu'il permet non seulement d'inscrire la langue dans une histoire déjà ancienne mais aussi dans une modernité attestant la vitalité du picard et sa possibilité de pouvoir dire la modernité avec ses propres mots.

2 — Cf. Martin F. (2015). *Espaces et lieux de la langue en Picardie au XXI^e siècle. Approche complexe de la structuration des répertoires linguistiques en situations ordinaires. Enquête en Picardie*, Thèse de doctorat, Université de Picardie Jules Verne, Amiens.

Conscients des limites que peut présenter la base dans son état actuel vis-à-vis de cette représentativité visée, nous œuvrons régulièrement pour affiner encore son contenu.

2. L'outil informatique au service du picard

2.1 La phase d'acquisition des textes du corpus

L'informatisation de la base PICARTEXT a reposé sur la mise en place d'une procédure rigoureuse permettant de traiter des données hétérogènes quant à leur format natif. Au-delà de la diversité typologique et géographique, la nature panchronique de la base a en effet réuni, de fait, des textes plus ou moins anciens et donc disponibles sur des supports différents. Le plus souvent proposés sur des supports papiers imprimés, les corpus récupérés ont par ailleurs également été proposés sous format manuscrit ou enfin numérique. L'une des premières difficultés du projet a donc été de gérer cet écueil important qui a mobilisé plusieurs personnes en charge de phases successives : passage à la reconnaissance optique de caractères (OCR)³, distribution des textes pour la mise en conformité de la sortie de l'OCR avec l'original, relecture du résultat de mise en conformité (offrant ainsi en quelque sorte une seconde relecture).

Le résultat de cette phase primordiale du projet est la mise à disposition de textes au format texte brut, sans aucune information typographique particulière (pas de hiérarchisation en titres, pas de gras, d'italiques, de surlignements, etc.) et donc manipulable par la plupart des outils informatiques relevant du traitement automatique des langues (TAL).

Le schéma proposé ci-dessous aidera le lecteur à se faire une idée un peu plus précise de la chaîne de traitements que les textes du corpus ont pu subir pour être ensuite interrogeables en ligne :

3 — Nous avons opté à l'origine pour la solution propriétaire et payante *OmniPage pro 10*.

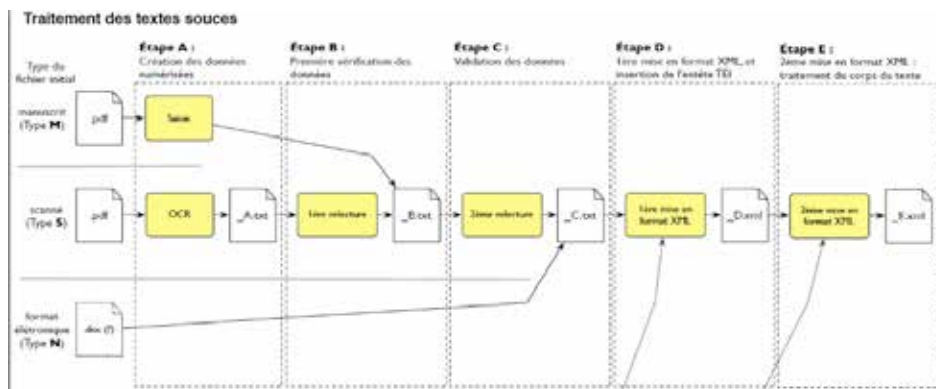


Figure 1. Chaîne de traitement des corpus présents dans PICARTEXT

2.2 La mise en place d'une plate-forme de gestion des textes

Cette première phase a bien entendu été accompagnée d'une inventurisation rigoureuse permettant de conserver toutes les informations documentant les textes de la base selon de nombreux critères tels que ceux de la représentativité décrite plus haut, mais aussi plus « basiques » comme ceux concernant l'auteur, la date, le titre, etc. Ces dernières informations se trouvent toutes stockées au sein de fiches bibliographiques informatiques dont les entêtes ont été converties au format XML. La génération automatique des fichiers bibliographiques XML permet notamment de pouvoir convertir ces données balisées en champs directement interrogeables via l'interface de consultation de PICARTEXT⁴, mais aussi de conférer à cette base une pérennité pour les décennies à venir, tant le format de codage retenu est incontournable aujourd'hui et s'est imposé au sein des plus grands projets de recherche dans le monde.

La mise en place d'un système de gestion permettant aux responsables du projet d'avoir un contrôle sur l'état du traitement de chaque fichier et l'avancement de la représentativité désirée de la base a été assurée successivement par deux post-doctorants⁵, recrutés pour leur profil idoine de linguiste et d'informaticien.

2.3 Un outil d'interrogation robuste et ambitieux

Désormais disponible en ligne par le biais d'un portail internet spécifique⁶, la base de données PICARTEXT dispose d'un module

4 — <https://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/>.

5 — Yayoi NAKAMURA-DELLOYE et Alain DAWSON.

6 — <https://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/>.

d'interrogation sophistiqué pouvant répondre tout aussi bien aux attentes du grand public qu'à celles des chercheurs. Nous reproduisons ci-dessous une copie d'écran de la page de recherche du projet :

Recherche d'un mot dans le corpus

Module expérimental de recherche de mots dans le corpus (concordancier).

Mot recherché (exemple : "tchair"):

Méthode de recherche :

- ☒ Chaîne littérale (ex.: trouve uniquement "tchair")
- ☐ Correspondance phonétique (ex.: trouve "tcherre", "tchère", "tcher...")
- ☐ Correspondance dialectale (ex.: trouve aussi "querre", "queure...")
- ☐ Expression rationnelle étendue : voir [cette page](#)

Lieu de référence des auteurs :

- ☒ Nord ☒ Pas-de-Calais
- ☒ Aisne ☒ Oise ☒ Somme
- ☒ Hainaut belge

Année de naissance des auteurs : Après Avant

Genres (plusieurs choix possibles) :

BO

Chanson

Chronique

Correspondance

Figure 2. Formulaire d'interrogation de la base PICARTEXT.

Les informations bibliographiques disponibles dans les fichiers XML évoqués plus haut permettent d'emblée, mais néanmoins sans aucune obligation, de pouvoir interroger la ressource en déterminant un corpus de recherche spécifique. Celui-ci peut être défini selon les différents paramètres que sont les zones géographiques de naissance des auteurs, l'empan temporel relatif aux dates de naissance de ces mêmes auteurs, ou encore le genre textuel auquel appartiennent les textes de la base. On peut ainsi imaginer la recherche d'un mot qui ne porte que sur des chansons écrites par des auteurs originaires du Nord et nés entre 1900 et 1950. La pré-sélection de ces choix avant de lancer une recherche donnerait le résultat schématisé ci-dessous :

Lieu de référence des auteurs :

- ☐ Nord ☐ Pas-de-Calais
☐ Aisne ☐ Oise ☒ Somme
☐ Hainaut belge

Année de naissance des auteurs : Après Avant

Genres (plusieurs choix possibles) :

Figure 3. Exemple de pré-sélection en fonction de critères préalables

À ces premiers critères viennent ensuite s'ajouter des méthodes d'interrogation et d'exploration du corpus davantage complexes et techniques.

Outre la recherche par « chaînes de caractères »⁷ et celle par « expressions rationnelles »⁸, l'utilisateur a en effet accès à des méthodes d'investigation plus élaborées, appuyées sur Dawson (2006) et utilisant l'approche théorique de Mc Carthy & Prince (1995). Il est ainsi possible d'effectuer des recherches :

- par « correspondance phonétique » : le mot est d'abord converti en sa représentation phonétique à l'aide d'un phonétiseur⁹. 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.

7 — Une recherche « plein texte » autrement dit.

8 — Il s'agit en fait d'expressions régulières.

9 — 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.

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. Nous reproduisons ci-dessous un exemple de résultat pour la recherche du mot « éfant » :

[illegible]

Figure 4. Concordances du mot « éfant » dans Picartext

À défaut de multiplier ici les exemples de recherches possibles dans l'état actuel de la base PICARTEXT, nous souhaitons achever cette présentation rapide des fonctionnalités de l'outil d'interrogation de la base en précisant que nous avons conscience qu'une étape nécessaire de la valorisation d'une telle ressource est celle de l'amélioration de son accès et de sa consultation pour le grand public. Pour être davantage et encore mieux consulté par l'ensemble de la communauté des picardisants mais aussi par les publics scolaires et les associations militantes, le portail d'accès à PICARTEXT doit être entière repensé dans son ergonomie, notamment en offrant plusieurs strates de consultation répondant aux attentes des spécialistes et non spécialistes. Il s'agit là d'un objectif faisant déjà l'objet de réflexions avancées qui, nous l'espérons, déboucheront sur une réalisation concrète à moyen terme.

3. Les avancées permises grâce à PICARTEXT

Objet hybride, PICARTEXT s'impose aujourd'hui non seulement comme un formidable outil de recherche permettant de fédérer autour d'un corpus, d'une ressource, les différents chercheurs et jeunes chercheurs travaillant sur la langue picarde, mais

aussi comme une ressource permettant, au delà de toutes les autres données éparses déjà disponible, de visibiliser le picard aux yeux du grand public.

En ce qui concerne sa dimension plus proprement scientifique, PICARTEXT, au regard des derniers travaux que nous avons à notre connaissance, constitue également un socle sur lequel de nouvelles initiatives peuvent venir s'articuler pour le futur.

Avec l'arrivée de la langue picarde dans le giron des langues régionales de France disposant, comme le corse ou l'occitan, de ressources textuelles de grande taille disponibles sous une forme électronique, de nouvelles initiatives peuvent en effet enfin être envisagées.

À l'état de « possible » hier, la constitution d'un dictionnaire globalisant pour la langue picarde – un peu selon le principe déjà adopté par Jean-Marie Brailon pour son *Dictionnaire général* – apparaît désormais comme une réalité plus que certaine. Plus largement, PICARTEXT a également ouvert la voie à de multiples processus de traitement automatique qui jusqu'alors n'avaient pu porter que sur des langues comme le français et l'anglais, langues richement dotées en données textuelles. C'est en ce sens que PICARTEXT a contribué à la conception et à la réussite du projet de recherche RESTAURE, au sein duquel les auteurs de cette contribution sont d'ailleurs très activement impliqués.

Avec RESTAURE, le picard – au même titre que l'occitan et l'alsacien qui sont les deux autres langues régionales impliquées – est désormais perçu comme une langue dont les spécificités linguistiques méritent non seulement d'être davantage décrites finement, mais s'affirme également comme une opportunité linguistique permettant de faire évoluer les outils du traitement automatique des langues.

À défaut de pouvoir nous étendre davantage sur cette initiative scientifique de grande ampleur¹⁰, nous renvoyons le lecteur au portail de présentation du projet¹¹ et achevons cette évocation en précisant qu'à la fin du projet RESTAURE, des réflexions abouties auront été engagées en faveur de la lemmatisation et de l'étiquetage morpho-syntaxique du picard... deux choses encore jamais réalisées à ce jour.

10 — Financée par l'Agence Nationale de la Recherche.

11 — <http://restaure.unistra.fr/>.

Conclusion

Avec la mise en place de la base PICARTEXT, nous avons pu doter la langue picarde de la plus grande ressource textuelle qui lui ait été consacrée à ce jour. De taille déjà conséquente, cette base est appelée à croître et à connaître de nouveaux développements pour répondre au mieux aux besoins de la communauté des chercheurs, lesquels, en se fédérant autour d'elle, vont encore affiner les investigations rendues possibles sur le picard.

Encore assez mal connue du grand-public, cette ressource doit peu à peu s'extraire de l'environnement scientifique dans lequel elle se trouve confinée, afin d'offrir aux picardophiles, picardophones et/ou picardisants un accès vers la découverte ou la redécouverte de leur langue. Cette phase de valorisation semble paradoxalement, peut-être, la plus difficile à réaliser tant les représentations sur les langues minoritaires ou minorées sont importantes partout dans le monde. Il ne fait nul doute que les chercheurs, les militants et les amoureux de la langue ne manqueront pas cette chance importante qui est donnée à la langue picarde de mieux rayonner.

Fanny MARTIN et Christophe REY
Université de Picardie Jules Verne

Références bibliographiques

- Braillon J.-M., 2001. *Dictionnaire général français-picard*, tome I, éditions F.I.P.Q.
- Dawson A., Éloy 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, Alain, 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. Soutenue à Toulouse le 14 décembre 2006. Lille : Atelier National de Reproduction des Thèses.
- 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.

- Éloy, J.-M., Martin, F., Rey, C., 2015. « PICARTEXT: Une ressource informatisée pour la langue picarde », *Actes de la Conférence Traitement Automatique des Langues Régionales de France et d'Europe (TALaRE'2015)*, 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2015), 22-25 juin 2015, Caen.
- Martin F., 2015. *Espaces et lieux de la langue en Picardie au XXI^e 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. Faithfulness and Reduplicative Identity, in J. Beckman, L. Walsh Dickey, S. Urbanczyk (éd.), *Papers in Optimality Theory*, U. of Massachusetts Occasional Papers in Linguistics 18, Amherst, Mass.: Graduate Linguistic Student Association, 249-384.