Appel à projets "Humanités numériques" MSHS 2017-2018

Bilan du Projet CRISTAL

Porteur de projet

Laboratoire FoReLLIS (EA3816)

Laboratoire MIMMOC (EA3812)

Rappel des objectifs

L'objectif principal de CRISTAL est d'impulser de nouvelles dynamiques sur des corpus scientifiques en réalisant des actions citées ci-dessous :

1. Numériser des ouvrages anciens/rares/manuscrits/contemporains permettant de produire des fac-similés donnant naissance à des corpus numériques.
2. Enrichir de champs descriptifs (aussi appelés étiquettes, métadonnées, index) décrivant les fac-similés.
3. Créer des ressources dérivées: Transcription OCR ou manuelle, traduction, étiquetage encodé, alignement.
4. Mettre en ligne des corpus numériques et leurs champs descriptifs sur des serveurs institutionnels (Université de Poitiers/Imédia/Service Commun de Documentation).
5. Archiver à long terme sur les serveurs d'Huma-Num/CINES les facs-similés (haute résolution) accompagnés de leurs champs descriptifs en respectant les normes d'**interopérabilité** indispensables au travail en réseau.
6. Explorer/Analyser/Traiter en vue de découvertes scientifiques dont l'utilisation de méthodes statistiques et algorithmiques.
7. Mettre en ligne des appareils critiques et des publications scientifiques basés sur l'étude de ces corpus.
8. Signaler ces corpus numériques à Isidore/Gallica.
9. Diffuser corpus et travaux scientifiques en étant acteur dans des consortiums (CAHIER, CORLI) et colloques.

Justification des financements MSHS

En soutenant ce projet, la MSHS de Poitiers a permis le recrutement de deux vacations d'appui aux développements des projets.

Recrutement IGE : Laurent Passion (141h)

Numérisation de 173 revues de presse Satirique espagnole de 1974-1978, permettant la création de 6954 images au format TIFF (139 Go), puis la production de 6954 images compressées au format JPEG (8 Go).

Un traitement massif de reconnaissance optique de caractères OCR pour produire 173 PDF (3 Go) contenant les images de la revue ainsi qu'un texte sélectionnable et interrogeable.

Un important travail archivistique a été réalisé pour décrire chaque revue et ses collaborateurs afin de faciliter sa découverte et son exploitation. Chaque revue est décrite en utilisant 13 champs normalisés DCTERMS.

Ces documents numériques PDF ont ensuite été versés sur le service Nakalona de la TGIR Huma-Num pour permettre leurs archivages, visibilités et explorations.

Site internet : https://porfavor.nakalona.fr

Recrutement IGE : Sophie Kraeber (141h)

Numérisation du dictionnaire de prononciation en langue anglaise Barclay de 1774.

Création de 1134 images au format TIFF (20 Go), puis la production de 1134 images compressées au format JPEG (1,7 Go).

Un traitement massif de reconnaissance optique de caractères OCR pour produire 1134 fichiers textes à corriger. En effet ce dictionnaire étant ancien, la reconnaissance optique automatique n'est pas suffisamment fiable pour permettre son exploitation directe.

C'est donc un important travail de curation de données sur les mots vedettes, la prononciation, la catégorie grammaticale et l'étymologie qui a été réalisée sur l'ensemble du document. Les définitions des mots n'ont pu être traitées dans le temps imparti. Cependant, les mots vedettes et leurs prononciations permettent d'initier des études diachroniques avec les autres dictionnaires traités dans le projet DicoDiachro.

L'intégralité des images, des re-transcritions et éditions des différents dictionnaires du projet DicoDiachro sont actuellement stockés et partagés (avec tous les membres du projet) sur le service Sharedocs de la TGIR Huma-Num pour faciliter le suivi et la maitrise d'ouvrage.

Bilan scientifique et technique des corpus

Corpus GRAFE multilingue écrit bidirectionnel – parallèle et comparable.

Acteurs : Raluca Nita, Michael Nauge, Joasha Boutault

Actions :

✓ Traitement des fichiers bilingues en vue de l'alignement :

- Vérification du toilettage suite à la numérisation et à l'océrisation
- Prise en compte de problématiques spécifiques au genre du corpus (ouvrages scientifiques en linguistique) en vue de l'alignement et suppression de certains éléments du corpus de départ (figures, tableaux, notes de bas de page). (*Cf* plus bas)
- Création de différentes versions électroniques des sources du corpus afin de garder trace de l'ensemble des données en vue d'une future exploitation des données concernées par la suppression :
    - Fichiers « Sources brutes » (fichiers comprenant les sources avec figures et notes de bas de page)
    - Fichiers « Sources pour alignement » (fichiers prêts pour l'alignement et ne comprenant pas les figures et notes de bas de page)

✓ Réflexion méthodologique :

Le traitement des données en vue de l'alignement a amené une réflexion méthodologique sur la nature des données entrant en ligne de compte dans l'homogénéité du corpus. Les éléments traités, à savoir figures, tableaux, notes de bas de page, posent problème lors de l'alignement (ne sont pas alignables).

Les notes de bas de page, constituent une source d'hétérogénéité quantitative et qualitative par rapport au du corpus :

- Elles peuvent en effet apparaître dans les originaux et être traduites, mais elles relèvent d'un genre différent par rapport au texte de départ (source d'hétérogénéité qualitative).
- Elles peuvent relever du traducteur ou de l'éditeur, et se retrouver respectivement dans l'original ou dans la traduction (source d'hétérogénéité qualitative).

Plus globalement, l'homogénéité quantitative des sources du corpus s'est constituée à partir du nombre de mots par la langue de départ. La prise en compte des notes de bas de page aurait amené leur inclusion dans le comptage et ainsi la réduction du texte « principal » - imbriquant, ce qui aurait à nouveau créé de l'hétérogénéité.

✓ Alignement :

Création d'une chaine de traitement semi-automatique d'alignement de corpus multilingue par le développement d'un logiciel workflowGRAFEtools.py utilisant Alinea assurant la compatibilité avec Paraconc pour l'analyse comparative.

Actuellement 18 fichiers sont alignés dans le sous-corpus GRAFE Linguistique Anglais-Français (9 par langue), pour un total de 6475 phrases comparables (142 925 mots dans les originaux, 155 354 mots dans les traductions), et 18 fichiers ont été préparés pour l'alignement dans le sous-corpus GRAFE Linguistique Français-Anglais.

Corpus en langue anglaise du fonds ancien Dubois

Acteurs : Elodie Peyrol, Susan Finding, Anne Sophie Traineau Durozoy, Nolwen Clement Huet, Michael Nauge

Actions :

Analyse des besoins.

Réunions de coordinations avec le SCD.

Numérisation, catalogage et indexation des ouvrages (réalisation SCD) :

- General Propositions relating to colonies [FD 195],
- A supplement to J. Massie's Brief Observations concerning the management of the war [FD 196],
- A reply to a pamphlet : called Observations on the Bank of England London : J. Whitlock.
- [FD 2268],

Alignement de vocabulaire Unimarc (SUDOC) vers DublinCore (Omeka)

Corpus “Por Favor”

Acteurs : Ludivine Thouverez, Laurent Passion, Anne-Sophie Pascal, Michael Nauge

Actions :

Ouverture d'un espace de travail collaboratif (sharedocs Huma-Num)

Mise en ligne d'une bibliothèque virtuelle (Nakalona Huma-Num)

Dépôt massif des documents (décrits) sur la bibliothèque virtuelle (151 revues soit 6146 images).

Site : https://porfavor.nakalona.fr/

Signalement au SUDOC (SCD) : http://www.sudoc.abes.fr/DB=2.1/SRCH?IKT=12&TRM=233991476

Réalisation d’une exposition "Chili Argentine 1974-1978, Les juntes militaires vues par les humoristes espagnols de Por Favor" du 7 au 22 décembre 2018 avec les étudiants en LLCER Espagnol. Ce travail collectif donnera lieu à un article scientifique à paraître dans la revue *Atlante*. De même, les dessins de l’humoriste Cesc feront l’objet d’un article dans le prochain numéro de la revue *Ridiculosa*.

Figure 2 Visuel pour l'exposition

Figure 1 Capture d'écran #1 de la bibliothèque virtuelle

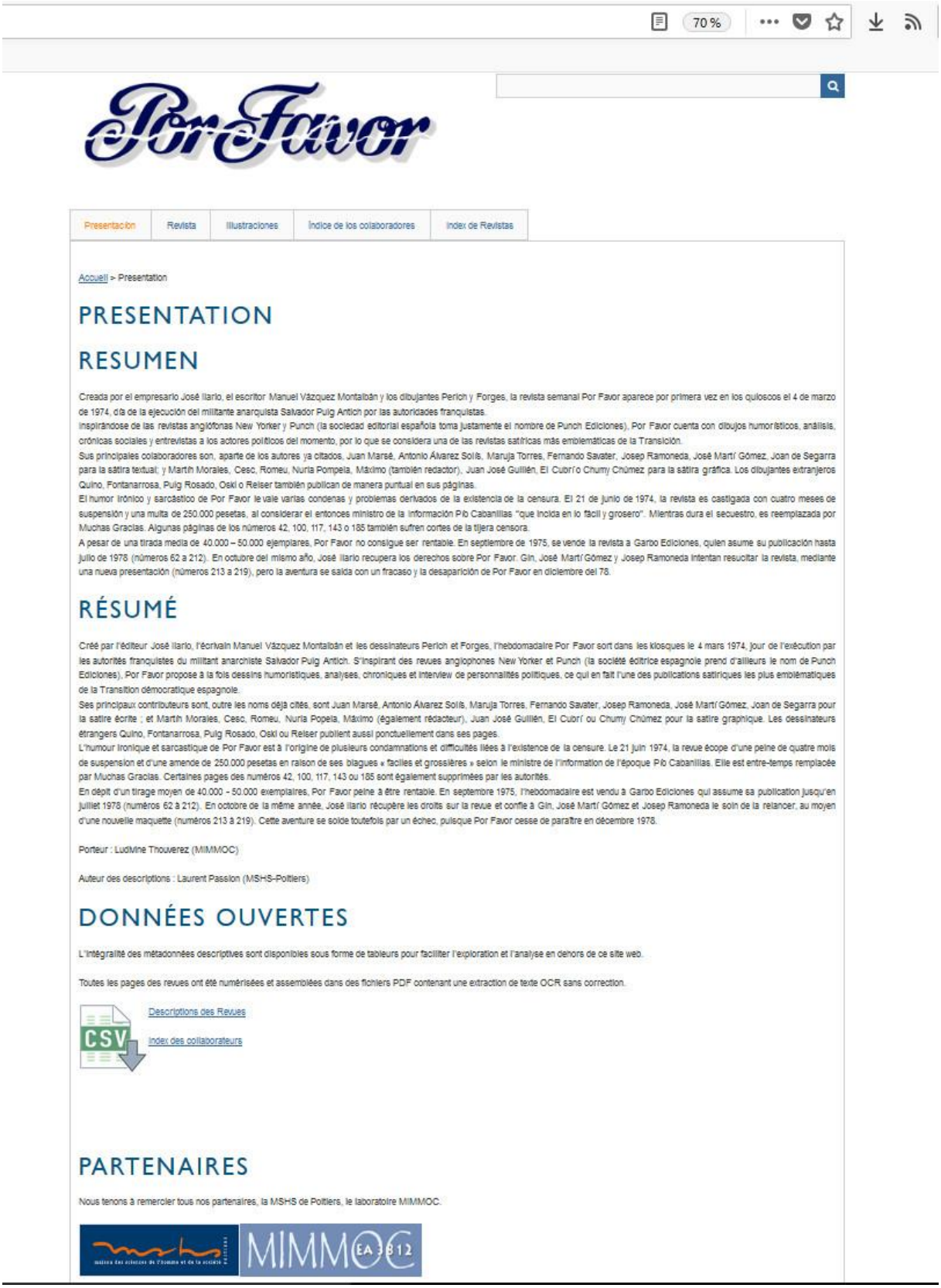

Figure 3 Capture d'écran #2 de la bibliothèque virtuelle

Corpus d'anglais diachronique “DicoDiachro”

Acteurs : Sylvie Hanote, Jean-Louis Duchet, Franck Zumstein (Université de Paris 7 - Denis Diderot), Nicolas Trapateau (Université de Nice - Sophia Antipolis), Jeremy Castanier (Université de Picardie - Jules Verne), Nicolas Videau, Nicolas Ballier (Université de Paris 7 - Denis Diderot), Sophie Kraeber, Michael Nauge

Actions :

Ouverture d’un espace de travail collaboratif (sharedocs Huma-Num).

Numérisation et retranscription du dictionnaire Barclay (1134 pages).

Utilisation de plusieurs logiciels OCR et automatisation de traitement pour l’optimisation de la retranscription.

Retranscription collaborative des 5 tomes du dictionnaire Wright (122/4829 pages) à partir d’un protocole commun discuté entre les membres du projet.

Prototypage de transformation des retranscriptions de Wright en format web.

Reprise de l'existant (7 dictionnaires) avec le développement de 7 scripts de transformations des fichiers d'entrées pour les convertir en données matricielles (point de pivot) en vue de recherches croisées sur plusieurs dictionnaires.

Curation de données (correction OCR + correction retranscription)

Premiers résultats d’exploration de données (génération de tableurs et graphiques) assistés par l’informatique (Jupyter Notebook)

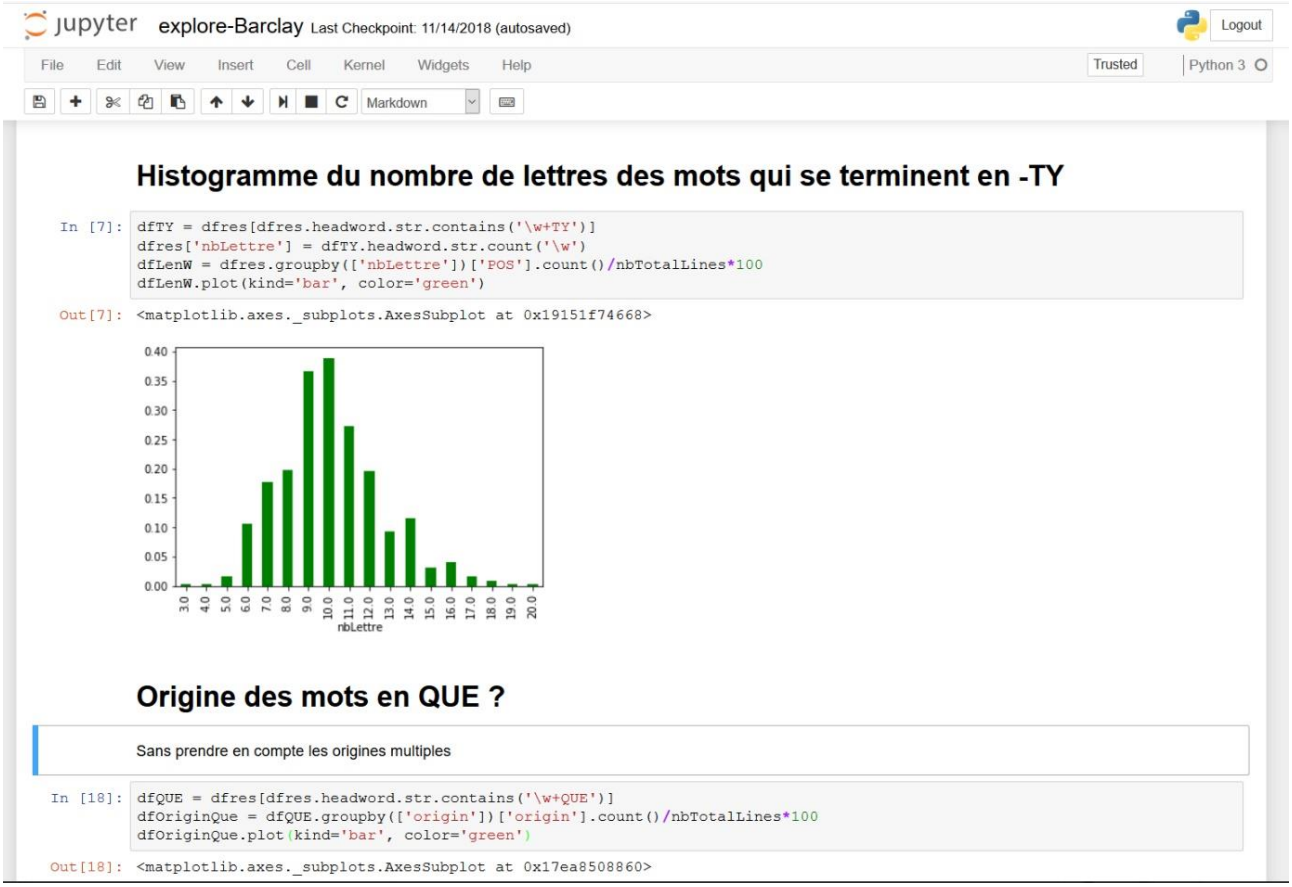

Figure 4 Exemple d'exploration de données

Référence des dictionnaires utilisés cette année :

- Bailey, Nathan, An Orthographical Dictionary, Shewing both the Orthography and the Orthoepia of the English Tongue, London: T. Cox, **1727**. [2e partie du volume 2, exemplaire de travail de Marie-Noëlle Antolin].

- Buchanan, James, An Essay Towards Establishing a Standard for an Elegant and Uniform Pronunciation of the English Language, throughout the British dominions, as practised by the most learned and polite speakers, London: E.N.C. Dilly, **1766**. [exemplaire de travail de Frédéric Duchesne.]

- Barclay, James, A Complete and Universal English Dictionary on a new plan, London: Richardson & Urquhart, **1774**, XXVI + 1120 p. [exemplaire de travail de Sophie Kraeber et Michael Nauge]

- Sheridan, Thomas. A General Dictionary of the English Language, one main object of which, is, to establish a plain and permanent standard of pronunciation, to which is prefixed a rhetorical grammar, London: J. Dodsley, C. Dilly & J. Wilkie **1780**, [24]+1029 p. [édition réalisée par Véronique Pouillon, Paris Diderot, 2017]

- Walker, John, A critical pronouncing dictionary, and expositor of the English language, 2nd ed., **1797**. [édition scannée en image de page, MSHS].

- Walker, John, A critical pronouncing dictionary, and expositor of the English language, 6th stereotype edition, **1809** [éditée réalisée par Nicolas Trapateau]. 26th stereotype edition, 1823; 29th stereotype edition, 1827.

- Jones, Stephen. A General Pronouncing and Explanatory Dictionary of the English Language, 3rd ed. London **1798**, 904 p.

- Wells, John C. Longman Dictionary of Pronunciation, Longman, **1990**, 2000, **2008**. [fichier électronique de Lionel Guierre pour l'éd. de 1990, CD-ROM éditeur pour l'édition de 2000, CD-ROM éditeur et fichier xml pour l'édition de 2008.]

Conclusion :

Le travail amorcé dans le cadre du projet DicoDiachro va permettre, à terme, d'avoir des données issues de différents dictionnaires de prononciation contemporains ou distants dans le temps qui permettront de faire des recherches en synchronie ou en diachronie (évolution de la prononciation des mots dans le temps) avec une base de données fiable et interopérable. Il s'agit de rendre le corpus disponible une fois le travail de saisie et de curation des données fini et de l'exploiter à des fins de diffusion scientifique (colloques, journée d'étude).