Constitution d'un corpus linguistique pour une analyse textuelle des discours à partir de la presse ancienne Le corpus *La Tribune Indochinoise*[1]

Thi Thanh Quyên Pham

Université de Franche-Comté – Besançon

Résumé
Dans un contexte de multiplication des sources pour la constitution des corpus linguistiques, la presse ancienne reste une des sources les plus sollicitées par les chercheurs en sciences humaines et sociales. Cependant cette source abondante dont la nature du support est très fragile, pose quelques difficultés pour l'établissement scientifique des données textuelles. La question se pose ainsi : quels sont les enjeux de la constitution du corpus à partir de la presse ancienne ?
Dans cet article, nous présentons la méthode de travail que nous avons utilisée face à ce genre de document en présentant la constitution du corpus *La Tribune Indochinoise*. Ainsi, l'objet de cet article est double. Il s'agira d'abord de montrer comment il est possible de corriger les défauts de la presse ancienne pour construire un corpus en mode plein-texte. Ensuite, nous présenterons la constitution du corpus *La Tribune Indochinoise*.

Mots clés : constitution du corpus, presse ancienne, philologie numérique, numérisation, océrisation (OCR), données textuelles.

1. Introduction

La constitution du corpus *La Tribune Indochinoise* fait partie des travaux de recherche entrepris dans le cadre d'une thèse de doctorat qui se situe dans le domaine des sciences du langage (Analyse du discours). Elle porte sur le discours du journal semi-hebdomadaire francophone *La Tribune Indigène*, puis *La Tribune Indochinoise*, paru en Indochine entre les deux guerres mondiales. La recherche consiste à éclairer le contexte socio-historique de l'activité des grands quotidiens coloniaux de l'ex-Indochine française, par l'exploration systématique de leurs matérialités discursives. Elle exige donc de fixer le panorama de ce journal de 1921 à 1940, et de recenser les archives disponibles afin d'en numériser des parties essentielles et / ou représentatives en mode texte, c'est-à-dire accessible à des recherches linguistiques. Nous prévoyons d'appliquer à ce corpus en constitution les outils conceptuels et logiciels de l'analyse textuelle des discours, afin de l'analyser sur les plans lexico-thématique, énonciatif et pragmatique. C'est une étude qui s'inscrit dans le champ des connaissances

[1] Sous la direction de Margareta KASTBERG - Maître de conférence de l'université de Franche-Comté

empiriques des médias et qui vise aussi à intéresser celui des sciences historiques et géopolitiques, ainsi que des sciences sociales en général, d'un point de vue interdisciplinaire.
La constitution de ce corpus nous fait entrer dans les problématiques de la *philologie numérique* (Viprey, 2005), domaine lié aux plus récents développements des sciences du langage, mais mobilisant d'autres champs de connaissances et de compétences, ceux des sciences des textes et de l'histoire des langues notamment, et exigeant des procédures très spécifiques. En dehors de ces problématiques principales, ce corpus nous présente en second lieu d'autres problématiques liées aux difficultés provenant des supports originaux des documents.

Nous voudrions premièrement présenter dans cet article la méthodologie et la technologie employées pour réaliser ce genre de corpus. Deuxièmement, nous ferons le point sur les difficultés rencontrées quand on travaille avec la presse ancienne en expliquant comment a été constitué ce corpus *La Tribune Indochinoise.* Enfin, nous terminerons par un bilan et une présentation des perspectives de ce travail de constitution de corpus.

2. La méthodologie de la constitution d'un corpus de la presse ancienne

La constitution de ce corpus commence par la numérisation. Cette dernière se fait souvent en deux grandes étapes : la numérisation en mode image et l'océrisation en mode plein-texte. D'une part, le mode image permet de préserver l'intégralité des propriétés d'un texte inscrit dans son *document*[2]. D'autre part, le mode plein-texte ouvre la voie à tout un champ d'études sur les textes. Au niveau méthodologique, la numérisation assure au texte une fidélité à son document, de la typographie à l'organisation topographique du texte. Au niveau théorique, la numérisation d'un document constitue l'ensemble des opérations de restitution du texte. Tel est le dessin général d'un établissement scientifique traitant des données textuelles à partir d'un document en support papier.

Ainsi, la numérisation est actuellement le mode le plus favorable de conservation et de diffusion des documents pour les nombreux avantages qu'elle apporte aux documents. Premièrement, elle rend un document plus maniable. Deuxièmement, grâce à elle, nous évitons une manipulation fréquente des documents fragiles. Ensuite, elle permet aussi de reconstituer des documents dégradés. *Grosso modo*, elle change radicalement le rapport entre le document et son support matériel, entre le document et son texte. Enfin, l'invention d'internet l'a favorisée davantage. L'internet facilite un accès plus rapide au document et à distance, quelque soit la localisation du lecteur par rapport au document. En particulier, il permet un accès aux outils informatiques tels que les outils d'analyse statistique textuelle (Lebart & Salem, 1994) permettant ainsi une exploration fine du texte.

Parmi ces outils, se trouve le Diatag-Astartex[3], développé par Jean-Marie Viprey, dans les laboratoires ATST (Archives, Textes, Sciences des Textes) et LASELDI (Laboratoire de Sémio-linguistique, Didactique, Informatique) de l'Université de Franche-Comté, dans le cadre du pôle ARCHIVE, BASES, CORPUS (ABC) de la Maison des Sciences de l'Homme de Franche-Comté. Inspiré du travail dans l'archive, depuis 1998 le pôle ABC et ses laboratoires développent un environnement spécifiquement dédié à l'étiquetage lexical,

[2] Le document est le support physique d'un texte. Un document est donc susceptible d'attester plusieurs textes et, réciproquement, un texte peut être transcrit sur plusieurs documents
[3] http://laseldi.univ-fcomte.fr/document/viprey/page_JMV.htm

morpho-flexionnel et potentiellement syntaxique autonome. Cet environnement s'intitule DiaTag (pour Dialogic Tagging) dont le principe consiste à alterner des phases automatiques et des phases ouvertes, interactives, conviviales par l'accès plénier aux ressources descriptives (dictionnaires, listes et automates contextuels). Il est étroitement lié à l'environnement Astartex dédié à l'exploration assistée (par exemple, requêtes par le lemme, la subordination et la catégorie grammaticale) et aux calculs statistiques.

Dans le cas de notre travail, l'environnement DiaTag est utilisé pour mettre le corpus en base textuelle exploitable qui sera à son tour explorée avec l'environnement Astartex.

2.1 La numérisation en mode image

Il s'agit d'un des modes de conservation et de diffusion le plus privilégié des bibliothèques actuelles. D'abord, la numérisation permet une simple visualisation sur écran des documents. Par ailleurs, il s'agit d'une technique rapide et peu coûteuse. Enfin, elle assure une meilleure protection pour les originaux.

Les choix du format de stockage et de la résolution varient d'une bibliothèque à l'autre. Néanmoins, la majorité des bibliothèques ont opté pour les formats JPEG, JFIF ou PDF, avec une résolution de 200 ou 300 dpi. Ces formats compressés et allégés associés à une résolution moyenne répondent aux objectifs de diffusion au grand public et de valorisation de documents fragilisés.

Dans le cas de la presse ancienne, la numérisation peut s'effectuer à partir de la collection papier du quotidien ou des microfilms. En l'occurrence, *La Tribune Indochinoise* a été numérisé en format JPEG à partir des microfilms disponibles à la BNF. Nous avons fait ce choix pour mieux protéger les originaux fragilisés par le temps. Cependant, les microfilms étaient d'une qualité médiocre et ont limité le résultat de la numérisation. Sur ce point, nous allons voir plus tard les difficultés engendrées pour le corpus *La Tribune Indochinoise*.

La numérisation en mode image constitue l'opération la plus simple du processus de l'établissement scientifique des données textuelles. Il s'agit simplement de scanner des documents papier ou des microfilms en document numérique – mode image. Une simple connaissance en numérisation et en matériel est suffisante pour accomplir proprement ce travail. C'est ainsi qu'aujourd'hui, les grandes institutions ont sous-traité cette mission à un spécialiste privé pour diminuer le coût de travail et aussi le temps nécessaire à la réalisation de leur banque de données numérique.

2.2 La restauration et l'océrisation[4] des documents anciens

Le passage du mode image au mode plein-texte passe obligatoirement par l'océrisation. Techniquement, il s'agit du traitement d'une image de texte sur laquelle intervient un logiciel de reconnaissance de caractères : le logiciel déchiffre les formes et les traduit en lettres. Une étape d'apprentissage est parfois nécessaire, c'est-à-dire qu'à chaque caractère non reconnu, il faut lui indiquer quelle est la lettre en question. Mais, malgré cet apprentissage, il existe toujours un taux d'erreur dans la reconnaissance de caractères, lié à la qualité du document initial, aux polices employées, aux notes et à la forme du texte... Pour que la reconnaissance

[4] Le terme océrisation dérive de l'abréviation OCR : Optical Character Recognition, c'est-à-dire en français : Reconnaissance optique des caractères

des caractères soit bonne, les fichiers d'images doivent être de qualité optimale. Il est donc important que la numérisation en mode image opte dès le début pour une bonne résolution. En même temps, une relecture humaine des travaux issus de l'océrisation afin de vérifier l'ensemble du texte est indispensable.

Dans le travail du corpus de *La Tribune Indochinoise*, l'océrisation a été réalisée à partir des images numérisées, obtenues à partir des microfilms. Sans prendre en compte les dégradations des originaux, les images des microfilms au niveau bitonal sont de qualité médiocre. Ceci entraine une mauvaise reconnaissance voire une perte des caractères du texte. Afin d'améliorer la qualité des images, nous sommes obligés de faire un traitement préalable avec *Bookrestorer*[5]. Parmi ses multiples fonctions, ce logiciel permet d'opérer un recadrage, d'effacer les petites tâches, de corriger l'éclairage, de redresser les lignes et d'atténuer l'effet de la courbure. Pour *La Tribune Indochinoise*, nous avons simplement fait la binarisation. Cette fonction de *Bookrestorer* permet d'optimiser la qualité du seuillage du document, en convertissant en noir et blanc les documents captés en niveaux de gris. À cet effet, deux paramètres sont manipulés : celui de filtrage ayant pour fonction l'élimination de certains parasites, et celui de profondeur, agissant sur l'épaisseur des contours des caractères. Après cette opération, le bruit sur le document est considérablement réduit, ce qui aide par la suite la reconnaissance des caractères. *Bookrestorer* permet de réaliser cette opération de deux manières, manuelle ou automatique. De la manière manuelle, la binarisation est faite page par page afin de trouver une correction juste pour chacune d'entre elles. La deuxième moitié de notre corpus a été faite de cette manière. Pour la première moitié, nous avons fait une binarisation automatique. C'est-à-dire, nous prenons par sondage un certain nombre d'images pour chaque année. A partir de ces images, une binarisation optimale est effectuée, sans être validée, les valeurs des paramètres sont notées. À la fin du lot, une moyenne de ces paramètres est calculée et appliquée à toutes les images de l'année.

Nous avons choisi ces deux modes d'opérations par gain de temps. Au début de notre travail, nous ignorions les défauts des originaux et ceux des images numériques obtenues à partir des microfilms. Le travail a donc été réalisé d'une manière habituelle en commençant par découper le journal en article et colonne par colonne avant de l'océriser. En moyenne, chaque numéro du journal est découpé en 60 articles. Ceci représente un nombre important d'articles par an. A la moitié de la collection, nous avons fait un test d'océrisation. Les résultats obtenus nous ont permis de réaliser que ce corpus avait besoin d'un traitement préalable. Seulement, il était hors de question de recommencer le travail de découpage qui nous a pris du temps. Ainsi, nous avons décidé de faire une binarisation automatique pour la première moitié et une binarisation manuelle page par page du journal pour le reste du corpus. Cette erreur de départ nous a permis de nous rendre compte qu'il était nécessaire d'effectuer un traitement préalable pour les travaux réalisés à partir de documents anciens. En voici un exemple dans les pages suivantes, les figures 1 et 2 sont représentatives du bruitage sur les feuillets traités sur lesquels nous constatons la marque très nette de certaines pliures sur l'image.

Vient ensuite l'océrisation à l'aide du logiciel *ABBYY FineReader* qui permet de convertir différents types de documents tels que les documents papiers scannés, les fichiers numériques vers des formats modifiables et exploitables. Il fonctionne sur le principe de reconnaissance des caractères permettant une transformation d'un fichier image en fichier texte. Dans un premier temps, le logiciel analyse la structure de l'image du document. Il divise la page en éléments distincts tels que les textes, les tableaux, les images... Les lignes sont définies en

[5] Un logiciel de numérisation créé par la société *I2S* permettant la consultation, la restauration, la sauvegarde et la publication des ouvrages numérisés dans leur intégralité

mots puis en caractères. Une fois que le caractère aura été isolé, il les compare avec un groupe de modèles d'images grâce auxquels des hypothèses sont avancées sur ce que représente le caractère. C'est sur cette base d'hypothèses que le logiciel analyse les différentes variantes des courbures des lignes en mots et de mots en caractères. Cette opération peut être faite d'une façon automatique ou manuelle. Pour l'océrisation manuelle, nous sélectionnons les zones de texte, une par une. Pour celle automatique, toutes les images sont importées et lues sans travail de sélection des zones de lecture. Pour *La Tribune Indochinoise*, l'océrisation a été faite d'une manière automatique puisque le journal avait été découpé en article. Le travail était donc simplifié.

Dans un deuxième temps, le logiciel va passer en revue toutes ces hypothèses, le programme prend la décision de vous livrer un texte qu'il pensera être conforme à l'image reconnue. En complément, *ABBYY FineReader* dispose de dictionnaires prenant en charge 38 langues. Cette option permet d'affiner l'analyse d'un niveau texte à un niveau mot. Grâce à la prise en charge du dictionnaire, le programme améliore la précision de la reconnaissance des documents et facilite les vérifications ultérieures de résultats.

Le résultat obtenu de *La Tribune Indochinoise* reste pourtant peu satisfaisant (près de 6,4% de bruits d'océrisation, autrement dit de caractères non reconnus). Mais pour ce genre de document, ce taux de bruits est inévitable. Ils viennent de la dégradation des originaux, de la qualité des papiers et de la technologie de l'imprimerie au début du 20ème siècle qui était caractérisée par l'irrégularité de l'impression. Une correction linguistique assistée par un logiciel spécialisé dans le domaine, puis une relecture humaine et une saisie manuelle sont absolument nécessaires pour obtenir une base exploitable.

TROISIÈME ANNÉE Nº 207 — LE NUMÉRO 0 $ 10 — MERCREDI 4 JANVIER 1928

LA TRIBUNE INDOCHINOISE

Organe officiel du Parti constitutionnaliste indochinois

Paraissant tous les Lundi, Mercredi et Vendredi

Directeur : BUI-QUANG-CHIÊU
Rédacteur en Chef : TRAN-VAN-TRI
Administrateur : NG.-KHAC-NUONG

ABONNEMENTS :
Un an 12 $, 6 mois 7 $ 00
Direction & Administration
72, rue Lagrandière
SAIGON
Téléphone nº 626

DÉPOT LÉGAL INDOCHINE Nº 4507

LA RENAISSANCE ANNAMITE

L'ÉDUCATION CIVIQUE, SOCIALE ET MÊME POLITIQUE PAR LA PRATIQUE DES SPORTS

A Monsieur Blanchard de la Brosse

On retrouverait dans la collection d'un journal qui s'appela *La Jeune Asie*, les conseils que nous donnions à la jeunesse indigène pour la pratique des sports. Les premiers Annamites qui s'essayèrent au jeu de foot-ball, dans les terrains vagues de Dakao, de Giadinh et de la Plaine des tombeaux, donnaient une pénible impression. Impression de malaise pour qui rêvait, comme nous, une revanche du sort pour ce peuple courbé sous un injuste destin et qui, déjà, sentait confusément, dans sa misère présente, comme une espérance lointaine de relèvement...

Etaient-ils assez pitoyables, dans leur cohue indisciplinée, ces joueurs qui fonçaient en troupeau sur le ballon, sans garder leur formation de combat et les postes confiés à leur vigilance de lutteurs improvisés ! L'inconfort débraillé de leur tenue faisait peine à voir. Avec cela, souvent, une absence fâcheuse de correction sportive qui eût effaré le plus débonnaire des arbitres. Bref, la *pagaye* dans tout son éclat discordant.

Vous le voyez, nous ne les ménageons pas. Mais attendez la fin...

Ils ne savaient pas, ils ne savaient rien des règles de ce jeu qu'ignoraient d'ailleurs les jeunes Français eux-mêmes, il n'y a pas quarante ans. Aucune âme généreuse ne s'était encore penchée sur cette masse confuse et désireuse de savoir, pour lui définir tout au moins ce que signifiait le *fair-play*, le respect chevaleresque de la règle. Pour ces non-initiés, la règle était une inutile entrave à un jeu qui simulait un combat, un siège, en un mot la guerre. Justement quelques Annamites en revenaient, de la guerre ; ils avaient vu combien le *fair-play*, ici le droit des gens, gênait peu les redoutables envahisseurs de notre sol. La ruse suffisait donc, la règle n'était pas de jeu...

⁂

L'éducation se fit lentement, comme qui dirait en première vitesse. Quelques militaires français attirés par certaines avances, donnèrent de bonne grâce les premiers conseils techniques. L'autorité militaire vit cela sans enthousiasme. Elle n'avait pas tout-à-fait tort : des incidents de parties pouvaient avoir des suites, judiciaires ou autres, auxquelles seraient mêlés des soldats français.

Mais entre temps arrivèrent de France les premiers *foot-balleurs* annamites. Ils furent les vrais initiateurs de la règle, les premiers définisseurs du *fair play*. Et l'Annamite, dans sa passion touchante de s'instruire, de s'évader de sa ténébreuse ignorance et de gagner les régions de lumière, mit toute sa ferveur à élégancier son jeu, à l'européaniser. Il y réussit. Déjà les deux camps en présence prenaient un aspect plus cohérent et plus discipliné. Les premiers matches eurent lieu. La merveilleuse souplesse de ces petits corps frêles auxquels il manquait encore du muscle et du poumon — mais que les nerfs soutenaient de leurs puissantes réserves de congénitale placidité — firent merveille aux premières exhibitions publiques. Dès ce moment, pour des sportifs *vrais*, observateurs au coup d'œil prompt et sûr, généreux et enthousiastes comme ils le sont tous, il y avait dans cette race un ressort admirable dont nous nous étions, pour notre part, toujours douté, ne serait-ce qu'en voyant de ces brins de femmes faire des lieues de marche avec leur *nho* accroché à leur flanc ou, dans le balancement de leur buste aux lignes tanagréennes, ramer debout, des heures durant, sur leur *ghe-luong* égayé du vase de fleurs en papier et des baguettes d'encens, dans la cadence berceuse d'une mélopée plaintive que l'écho des rives du fleuve se répétait dans les nuits claires...

⁂

Avant que des équipes ne se formassent et que des matches réguliers ne s'organisassent, un Annamite qui fut un précurseur plein de foi et de charme, avait conquis déjà ses compatriotes par sa maîtrise sportive, d'une irréprochable correction. C'est ce gracieux Nguyên-phu-Khai, fondateur avec Bùi-quang-Chiêu de la vieille *Tribune indigène* et dont l'enthousiasme effréné eut parfois le tort de ne point assez pressentir l'obstacle... et la malignité humaine.

Il est certain que la maîtrise de Nguyên-phu-Khai, bien qu'elle se prodiguât peu, aura hâté l'éducation sportive de ses compatriotes et en cela leur aura rendu un éminent service.

⁂

Car pour un peuple en instance d'autonomie nationale, le sport — le sport collectif surtout, plus accessible à la masse — est une merveilleuse école de tactique, de sang-froid, de discipline (ça, c'est pour mes amis annamites) et aussi de **probité dans le choix des moyens** (ça, c'est pour le gouvernement de Cochinchine qui devrait, s'il est clairvoyant, ne point refuser au Cercle Sportif annamite les moyens financiers de développement qu'il prodigua intelligemment au Cercle Sportif européen, à ses débuts).

Pas plus qu'une équipe de foot-ball ne demanderait le gain d'un match à une supercherie, un peuple devenu fervent de cette religion du sport, n'accepterait d'exposer le sort de ses aspirations nationales à des tactiques improvisées et aventureuses, où la voix de ses chefs élus se perdrait dans les hurlements d'extrémistes provocateurs.

La marche de l'humanité vers la lumière est lente et douloureuse. Celle d'un peuple vers la liberté l'est aussi. Il faut savoir attendre son heure — nous ne disons pas l'heure du colonialisme dominateur, *qui ne viendrait jamais* —. L'heure du peuple annamite sera celle où ses chefs élus, dans leur inlassable effort, aura, par une propagande loyale, à ciel ouvert, conquis le cœur des élites françaises.

Cette heure est proche.

Jacques DANLOR.

Au profit des inondés de Thanh-hoa

P. V. du 3 Janvier 1928
Matches du 2 Janvier 1928
GROUPE A
E. Giadinh bat Cholon par 1 à 0 (homologué)
GROUPE B
Thudaumot bat Victoria par 3 à 1 (homologué)
Matches à jouer le 3 Janvier 1928
3 h. 15 à 4 h. 15
Sporting contre Télépost
Arbitre : M. Nhiêu.
4 h. 20 à 5 h. 20
Dakao contre E. Saigon
Arbitre : M. Qui Thomas.

Au profit des inondés de Thanh hoa

Recettes de la journée du 1er Janvier :	195 $ 50
Dépenses diverses	7 00
Reste net	188 $ 50

La Commission décide d'envoyer la somme de 188 piastres à M. le Résident de France à Thanh-Hoa pour être employée aux œuvres de secours des sinistrés de la province.

La Commission adresse ses chaleureux remerciements à M.M. Dauverchain et Triêu-van-Yên pour les cadeaux de Cognac Jules Robin et de pochettes destinés à nos équipiers qui ont participé à l'œuvre de bienfaisance.

Le retour de M. Nguyên-thê-Truyên

Nous apprenons que M. Nguyên-thê-Truyên qui a présenté avec intelligence et courage les revendications annamites en France, sera de retour parmi nous le vendredi 6 janvier.

M. Nguyên-thê-Truyên est accompagné de sa femme, française d'origine, et de ses enfants.

La « Tribune Indochinoise » est heureuse de présenter à notre vaillant compatriote et à sa famille ses meilleurs souhaits de bienvenue.

ÉDITORIAL

Les couvents maçonniques et la colonisation

Notre ami Jacques DANLOR, dans l'Editorial du 21 dernier émettait ici même quelques réflexions sur les récentes décisions de la Franc-Maçonnerie française à l'endroit de la colonisation.

Il disait notamment : « Ça ne casse rien... Si tout en proclamant que l'émancipation complète est le terme obligatoire de nos bonnes œuvres coloniales, on ne précise pas l'heure de cette émancipation.... il est certain que la question restera lettre morte. »

J'en demande pardon à notre ami, mais je ne suis pas tout à fait de son avis.

Ce n'est pas seulement le Grand-Orient *mais aussi la* Grande Loge de France *et le* Droit Humain, *c'est-à-dire toute la Maçonnerie française, qui a pris cette importante décision.*

Pourquoi l'a-t-elle prise ? Quelle en est la portée pratique ?

Le problème colonial est, pour l'heure, celui qui affecte le plus dangereusement la paix du monde. Aucun peuple colonisé — même certaines peuplades nègres qui ne constituent pas à proprement parler des peuples — n'accepte plus une domination étrangère, ne consent à refouler plus longtemps ses légitimes aspirations nationales. Il suffit de [illegible] une fois du « paite indochinois » où nous sommes enterrés vivants pour apercevoir ce fait patent qui saute aux yeux. Le monde est plein de ce fait.

Il est possible que l'heure de la délivrance ne sonne pas encore, mais la « résignation » est finie.

La Franc-Maçonnerie qui compte pour quelque chose dans le monde — et la Franc-Maçonnerie française a ici son mot à dire — se doit de marquer le point, de se saisir de la question et d'en indiquer la réponse.

Et elle dit : « Il faut que le but de la colonisation soit sans conteste l'indépendance de la colonie ». *Et il n'est pas sans utilité de fixer et proclamer une loi.*

Lorsque le droit pénal décide que le fait de s'approprier le bien d'autrui est un délit punissable, lorsque le droit international contemporain dit qu'un Etat agresseur se rend coupable du crime de guerre, on n'a pas fait qu'une déclaration platonique : on a enregistré une manifestation de la conscience universelle. On a matérialisé une loi naturelle pour la sanctionner.

Une décision de la Franc-Maçonnerie n'est pas un discours officiel. La Franc-Maçonnerie constate aujourd'hui qu'est contraire à l'évolution humaine, à la nature des choses, tout travail colonial qui ne tend pas à l'indépendance des peuples colonisés. Elle condamne tous ceux qui se servent de la colonie pour d'autres fins. Elle trace donc la conduite à suivre aux hommes libres et de bonne foi.

Cela est un acte, et un acte d'une haute portée humanitaire.

Et la Franc-Maçonnerie n'est pas seule à penser et à agir ainsi.

Ici, en Indochine, tous les Annamites croient fermement que leur pays recouvrira son indépendance un jour. Nombreux sont les Français qui pensent de même, l'idée semée commence à germer.

Qu'ils travaillent ensemble à réaliser rapidement ce but hautement humain.

D. V. G.

NHATRANG

L'affaire Vuong gia Ngai et Vuong gia Bât (1)

La félonie d'un mandarin et l'arbitraire d'un résident

Nous avons promis à nos lecteurs de les tenir au courant de cette triste affaire où l'arbitraire de certains agents du gouvernement du Protectorat et la vilenie de ses « dignes serviteurs » se sont joints pour montrer le peu de cas qu'il est fait de la liberté humaine sous la « tutelle bienfaisante de la France libérale ».

Nous sommes heureux de pouvoir aujourd'hui leur mettre sous les yeux le récit détaillé de *l'affaire des frères Vuong gia Ngai et Vuong gia Bât condamnés en* **1927** *l'un à 3 ans l'autre à 2 ans de prison* **pour avoir vendu la « Biographie de Tchang kai Chek ».**

M.M. Vuong gia Ngai et Vuong gia Bât avaient établi à Nha-trang une librairie dénommée « *Mông luong thu quan* ». Les noms des brochures interdites leur *étaient toujours* signifiés par le huyên et le commissaire de la localité qui venaient régulièrement donner lecture aux libraires des arrêtés d'interdiction et prendre possession des livres proscrits.

Cette procédure dérangeait naturellement les huyên et commissaire et les libraires leur suggérèrent l'idée que le Ban-ta T.T.D., une de leurs connaissances, pourrait les remplacer dans l'exercice de cette... corvée. Il fut entendu que désormais le Ban-ta leur signifierait les arrêtés d'interdiction.

Tout alla bien pendant un certain temps.

Puis, le 25 septembre 1927, ce dernier qui venait tous les jours prendre... [illegible] chez les libraires, leur dit en substance en riant : « *J'ai entendu dire vaguement* qu'il vient de paraître un arrêté interdisant la vente de la « *Biographie de Tchang kai Chek* » Est-ce que, par hasard, l'ami Bât n'aurait pas ce livre en magasin ?

— « J'en ai même deux de biographies de Tchang kai Chek. Est-ce qu'elles sont toutes les deux visées ? »

— « Non, il y en a seulement une d'interdite mais j'ignore laquelle. Pourriez-vous me donner un exemplaire de n'importe laquelle je verrai quelle est celle qui est interdite par l'arrêté et je vous aviserai pour que vous puissiez livrer les brochures à l'Administration ou les renvoyer à l'éditeur, car du moment qu'elles sont proscrites il est inutile de chercher à les vendre. »

Trompé par le ton amical du Ban-ta qui d'ailleurs, nous l'avons dit, était une de leurs vieilles connaissances (la femme du Ban-ta est même parente de la feue femme de M. Vuong gia Bât), M. Bât lui remit un exemplaire de la « *Biographie de Tchang kai Chek.* »

Le Ban-ta s'en alla donc avec la brochure. Il revint le lendemain 26 septembre vers les 8 heures du matin et dit aux libraires : « Le livre que vous m'avez remis hier, je l'ai mis en sûreté. J'en ai parlé à l'administrateur-adjoint qui m'a dit de vous informer d'avoir à m'en livrer tous les exemplaires que vous auriez en magasin. Cela vous épargnera les frais d'envoi par poste. » — M. Bât répondit : « Il m'a toujours été donné reçu des livres que j'ai livrés afin que je puisse m'en prévaloir auprès des éditeurs et obtenir remise de leur prix. Est-ce que M. l'administrateur-adjoint ne vous a rien donné comme ordre à ce sujet ? » — « L'adminitrateur adjoint m'a chargé de vous dire ce que je vous ai dit. Maintenant si vous préférez opérer régulièrement et aller liver les brochures interdites à la Résidence, vous êtes libre. »

Pensant que le Ban tá est une notabilité de la province et que d'autre part du moment qu'il avait l'ordre de l'administrateur-adjoint de prendre possession des livres, il valait autant les lui livrer, les deux libraires lui firent remettre les 21 Biographies qui leur restaient. Avec celui que le Ban tá avait déjà pris la veille, cela faisait 22 brochures dont ils demandèrent reçu.

— « Pour le reçu, je dois voir d'abord l'administrateur adjoint, dit le Ban tá ».

Le lendemain matin, il revint en effet et dit aux libraires : « C'est entendu, je vais vous établir le reçu. »

Et les deux frères lui ayant apporté de l'encre et du papier, il écrivit :

« Reçu de M. Ngai 22 brochures Tuong-gioi-Thach, livres interdits par l'Administration.

« Nhatrang, le 26 septembre 1927. »

(*à suivre*)

---

(1) Voir la « Tribune Indochinoise » du mercredi 21 décembre 1927.

En lisant la presse parisienne

L'Affaire des concessions du Darlac

Les lecteurs de la « *Tribune Indochinoise* » se souviennent sans doute de l'affaire des concessions au Darlac (région moïe en Annam central) accordées à un groupe de financiers par M. Alexandre Varenne, Gouverneur général titulaire, affaire contre laquelle « *La Liberté* » avait mené, à Paris, une violente campagne.

Ce scandale a également soulevé un important débat à la Chambre des députés, le 18 mars 1927, à l'issue duquel M. R. Poincaré, d'accord avec le Conseil des ministres, délégua en Indochine une Commission d'enquête.

Présidée par un ancien Directeur de la Sûreté générale, le Conseiller d'Etat Richard, cette Commission, composée de M. M. Blondeau et de la Randière, Inspecteur des Finances, est arrivée à Saigon le 3 mai suivant par le *d'Artagnan*.

Après trois mois d'investigations tant au Darlac qu'à Hué et Hanoï, tenues dans le plus grand secret, la Commission est rentrée, en France, en août dernier, pour soumettre son rapport au Président du Conseil.

Quatre mois se sont écoulés depuis sans qu'aucune nouvelle officielle sérieuse ne nous fusse parvenue, sauf deux laconiques « *communiqués* » de l'A. r. i. p. (*Agence radiotélégraphique Indochine Pacifique*) annonçant que la Commission « relevait quelques irrégularités dans l'octroi des concessions, mais sans admettre qu'il y ait eu collusion de fonctionnaires ».

Quant à M. Alexandre Varenne, mis en cause par ses adversaires politiques, tant dans la presse que devant le Parlement, il a, dès son arrivée à Marseille, en novembre dernier, déclaré aux journalistes venus l'interviewer, qu'il réclamerait la publication du Rapport en question.

Voilà tout ce que nous savons sur un scandale dont les échos ont retenti jusqu'à la Chambre en d'orageuses interpellations, et dans lequel s'affrontèrent des politiques en vue, des journalistes de marque, de hauts fonctionnaires.

Cependant, la presse parisienne, arrivée par le dernier courrier, soulève un peu le coin du voile qui cache plus d'un mystère.

En effet, dans le « *Temps* » du 20 novembre 1927, en 3e page, sous le titre : *L'enquête sur les concessions du Darlac*, en un filet d'une cinquantaine de lignes qui passerait sans doute inaperçu, tant la lecture de cet organe de la haute finance est austère, voire indigeste, nous lisons :

L'enquête sur les concessions du Darlac

La commission des Colonies a entendu hier M. Léon Perrier, ministre des Colonies, à qui des précisions avaient été demandées sur les résultats de l'enquête instituée dans le Darlac au sujet de l'octroi de concessions à une société privée.

Aucun communiqué n'a été remis à la Presse à l'issue de la réunion.

D'après les indications données par plusieurs assistants, le ministre a déclaré que la commission d'enquête présidée par M. Richard, avait remis un rapport qui a été transmis au comité du contentieux du ministère dont il ne pouvait, par conséquent, donner lecture.

Invité à faire connaître, néanmoins, les conclusions de ce rapport, M. Perrier a déclaré que ce document relevait certaines irrégularités dans l'octroi des concessions sans admettre, cependant, qu'il y ait eu collusion de fonctionnaires.

Interrogé ensuite sur les motifs de la démission de l'un des membres de la commission d'enquête, le ministre a répondu que cette démission n'avait pas été portée officiellement à sa connaissance, mais qu'il avait constaté au bas du rapport l'absence de la signature d'un inspecteur des finances qui faisait partie de la commission d'enquête.

Plusieurs membres de la commission ont soutenu alors que, d'après leurs renseignements personnels, cette abstention, sinon cette démission, avait été motivée par une sérieuse divergence de vues entre cet inspecteur des finances et le président de la commission d'enquête, principal auteur du rapport, en particulier au sujet des conclusions concernant le rôle joué dans cette affaire par les fonctionnaires.

La commission des colonies n'a donné aucune suite à l'audition du ministre ; il a été simplement pris acte de ses déclarations par certains des membres qui avaient, il y a plusieurs mois, attiré l'attention du Parlement et du Gouvernement sur cette affaire.

Dix jours après, le « *Quotidien* » organe du Cartel des gauches, a publié, en 1ère page, ***sans aucun commentaire***, la lettre reproduite ci-dessous, qu'un membre de cette Commission d'enquête, M. de la Renaudière, a adressée à M. Perrier, Ministre des colonies :

Compatriotes ! !

L'inondation à Thanh-hoa et à Nghê-an a détruit 62.542 hectares de rizières, causé 1.876.680 $ 00 de dégâts, privé de nourriture pendant 6 mois 156.355 de nos frères de race, démoli 5.705 maisons, jeté sur le pavé 23.525 hommes, femmes, vieillards et enfants, (chiffres officiels).

Une souscription en leur faveur est ouverte au bureau de la Tribune Indochinoise *nº 72 rue Lagrandière.*

Les généreux donateurs sont instamment priés d'envoyer les mandats avec leurs noms et domicile à notre administrateur M. Nguyên khac Nuong.

Compatriotes, secourez vos frères qui sont dans la détresse !

LA TRIBUNE INDOCHINOISE

N. B. — Mentionner sur les mandats le mot : **Souscription.**

Figure 1 : page 1 du 4 janvier 1928 de La Tribune Indochinoise

3. Constitution du corpus *La Tribune Indochinoise* : l'enjeu

Le premier objectif lors de l'établissement de notre corpus était de constituer un corpus en vue d'utilisations standards et spécifiques en analyse du discours. Le corpus est donc constitué avec le souci de préserver toutes les caractéristiques pertinentes de la source papier (emplacement topographique des articles dans le plan du journal, données typographiques, etc.), et d'éviter la moindre perte de texte. Or le résultat obtenu pour ce corpus est médiocre. Pour nous rendre mieux compte de ces résultats peu satisfaisants, nous allons comparer ici les résultats de deux bases textuelles ayant le même parcours de

constitution : *Le Petit Comtois*[6] fait dans notre laboratoire en 2006 et *La Tribune Indochinoise*. Ci-dessous, une image du *Petit Comtois*.

20 centimes

LE PETIT COMTOIS

JOURNAL RÉPUBLICAIN DÉMOCRATIQUE QUOTIDIEN

Rédaction et Administration : 20, Rue Gambetta, BESANÇON

20 centimes

MERCREDI 4 JANVIER 1928

OPINIONS

Les problèmes franco-allemands

LES NATIONALISTES ALLEMANDS PARTENT EN GUERRE CONTRE L'OCCUPATION

Le Livre de Prières

Les Soviets expirent, un jour, avec les tsars !

Un incident italo-autrichien

UN CARGO FRANÇAIS A DISPARU

Un lynchage en tranche les doigts

Un rez-de-mort s'abat sur la côte méditerranéenne

Il n'y a pas d'incidents en Syrie

C'est une réalité malchance

La mission Proust-Gama au départ de Perpignan abandonne un de ses appareils

LA GUERRE NORS LA LOI

Les Etats-Unis rejettent la proposition Briand

Avant la visite de Lindbergh

De sanglants combats mettent aux prises nicaraguayens et américains

M. MUSSOLINI VA PROTESTER CONTRE L'ASSASSINAT DU VICE-CONSUL COZZIO

Un nouveau plan réglant les dettes et les réparations ?

Après la neige, les inondations

Informations diverses

Il avait injurié un maître

Les pourparlers économiques franco-belges

Le pétrolier Italien sombre

LES TROUBLES AU MEXIQUE

INFORMATIONS FINANCIÈRES

Figure 2 : page 1 du 04 janvier 1928

Les deux bases contiennent un détail différent. *Le Petit Comtois* est numérisé à partir de document papier en TIFF (*Tagged Imaged Filed Format*), format flexible non-compressé à 301 dpi, tandis que *La Tribune Indochinoise* est faite à partir de microfilms au format JPEG. Cette différence influence fortement les résultats issus de l'océrisation des deux bases textuelles. Sur la figure 1, la lisibilité des caractères est moyenne. Certaines lignes de caractères sont même effacées. Dans l'ensemble, la collection se présente dans cet état. Nous pouvons également observer sur la figure n°3 que certains passages sont compromis par la transparence du *verso* sur le *recto*.

[6] La base de données du Petit Comtois est accessible à l'adresse http://laseldi.univ-fcomte.fr/petit-comtois/accueil.php

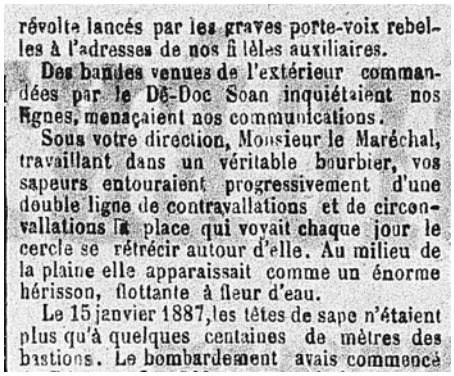
révolte lancés par les graves porte-voix rebelles à l'adresses de nos fi lèles auxiliaires.
Des bandes venues de l'extérieur commandées par le Dê-Doc Soan inquiétaient nos lignes, menaçaient nos communications.
Sous votre direction, Monsieur le Maréchal, travaillant dans un véritable bourbier, vos sapeurs entouraient progressivement d'une double ligne de contravallations et de circonvallations la place qui voyait chaque jour le cercle se rétrécir autour d'elle. Au milieu de la plaine elle apparaissait comme un énorme hérisson, flottante à fleur d'eau.
Le 15 janvier 1887, les têtes de sape n'étaient plus qu'à quelques centaines de mètres des bastions. Le bombardement avais commencé

Figure 3 : passage compromis par la transparence du recto sur le verso

Tandis que ce problème n'est pas souvent rencontré dans *Le Petit Comtois*.

Considérons à présent la qualité du mode texte obtenu pour les deux journaux. Pour pouvoir comparer, nous avons effectué l'océrisation de la dernière colonne de chaque image. Nous présentons en parallèle ci-dessous ces fragments des pages sélectionnées avec son équivalent plein texte, à gauche le résultat obtenu pour *La Tribune Indochinoise* et à droite celui pour *Le Petit Comtois*. Nous n'avons procédé à aucune modification de mise en forme et de correction des données textuelles.

Nijvis apprenons que M. Nguyôn-thè-Truyèn qui a présenté avec intel-ligenee "et courage les revendii-ations aiinauiites en France, sera de retour parmi nous le vendreili 6 janvier.
M. Nguyêri-ihô-Trnyén est ai'c-oni pagnê de sa femme, i'rançiise d'ori gine, et de .<e.s l'iifjnrs,
La « Tribune Indûchinoiso » est licurease de prt'isentc.i- à noire vaillant c.irapairiûtc et à sa famille ses meilleurs souhaits de bienvenue,

Figure 4 : OCR de la Tribune Indochinoise

C'est ce point de vue que le ministre de l'intérieur a soutenu à la Chambre des Communes, et il a vu se rallier à sa thèse des socialistes, des travoillislcs, des libéraux et d«s conservateurs, fous réconciliés dans la même crainte de voir le pontife romain établir son contrôle sur l'Eglise.

Figure 5 : OCR du Petit Comtois

Visiblement, sur la partie gauche, certaines erreurs sont imputables à la faible reconnaissance de l'accentuation des formes graphiques ; d'autres, à la suppression d'espaces figurant pourtant sur le support-papier, tandis qu'une partie des taches du document d'origine sont manifestement analysées comme de la ponctuation. Un autre type d'erreurs concerne la reconnaissance des caractères mêmes, aboutissant à la non-reconnaissance de mots entiers. Souvent, l'océrisation reconnaît un « e » pour un « o » ou un « c » ou le contraire, un « rn » pour un « m », un « f », un « i » pour un « t », un « a » pour un « n », un « l » pour un « i »…. Tandis que sur la partie droite, les erreurs sont limitées.

Ainsi, la difficulté de ce genre de travail est de rattraper toutes les pertes ou les mauvaises reconnaissances de caractères du texte. Pour des bases composées de centaines de milliers d'occurrences, cet inconvénient ne pose pas de problème conséquent. Mais quand un corpus atteint des millions, voire des milliards de mots, il est nécessaire de trouver un moyen efficace. Pour la constitution de notre corpus, nous avons adopté la solution semi-automatique avec le logiciel Diatag-Astartex qui permet une correction automatique des erreurs récurrentes, et une ressaisie du texte manuelle pour les passages totalement corrompus, tout en restant ergonomique.

La constitution de cette base de 4,5 millions de mots a nécessité 500 heures de restauration et d'océrisation et presque le quadruple pour corriger les erreurs issues de l'océrisation. Nous estimons que c'est un temps nécessaire et raisonnable que seul un laboratoire d'expérimentations méthodologiques et techniques peut réaliser avec pour objectif l'obtention d'un corpus pertinent et exploitable.

4. Correction linguistique

La correction sur *La Tribune Indochinoise* est effectuée en deux étapes. La première étape consiste à identifier et corriger les formes non reconnues. Quant à la deuxième étape, il s'agit de ressaisir les passages totalement corrompus.

4.1 L'identification et la correction des formes non reconnues issues de l'océrisation

Cette étape se fait en deux temps : l'identification des formes individuelles, le recensement et la validation des fautes corrigibles.

Les formes non reconnues provenant de mots étrangers, de noms propres, de néologismes…, ou bien les « bruits » de l'océrisation représentent 6,4% des items du corpus. L'identification et la correction de ces formes représentent un travail colossal. Face à une telle difficulté, la correction doit être ordonnée d'une manière ergonomique afin de ne pas y consacrer trop de temps et trop de ressources humaines. Une correction strictement manuelle (autrement dit reprendre mot à mot sur le clavier) demande beaucoup de temps. La correction entièrement automatique ne convient pas non plus à cette base car elle laisse une part conséquente d'erreurs. Donc, ces deux manières de correction sont exclues. Il nous reste la correction semi-automatique. Le principe de cette méthode est d'appliquer en premier lieu les dictionnaires[89] intégrés dans le logiciel sur le corpus pour repérer les formes courantes et les corriger tout de suite.

[89] Ce dictionnaire se compose d'un dictionnaire construit à partir du trésor de la langue française, d'un dictionnaire des langues étrangères : latin, anglais, russe, espagnol, italien et celui des noms propres courants

Dans un deuxième temps, les formes non reconnues du dictionnaire et les bruits issus de l'océrisation seront traités via l'interface de Diatag. L'opération est composée de deux phases. La première consiste à corriger les formes non reconnues lors de la correction automatique et ensuite à les intégrer au *rescor*. Le dossier *rescor* est constitué de fichiers où des formes erreurs sont accompagnées de leurs formes corrigées. Chaque fichier est représenté par un chiffre et une lettre. Le chiffre est le nombre de caractères qui composent la forme inconnue (pas sa forme correcte). La lettre est son caractère initial. Par exemple, le dossier *7a* contient toutes les formes contenant 7 caractères et commençant par la lettre a, tels que : ahemand (allemand), actioks (actions), arrivéo (arrivée)…. La deuxième phase consiste à préparer les entrées aux dictionnaires (dictionnaire du système et dictionnaire du corpus[90]) des nouvelles formes. L'opérateur dispose d'un dictionnaire général, d'un dictionnaire particulier, d'un dictionnaire des noms propres particuliers, d'un dictionnaire des noms propres généraux et d'un pour les mots d'origine étrangère.

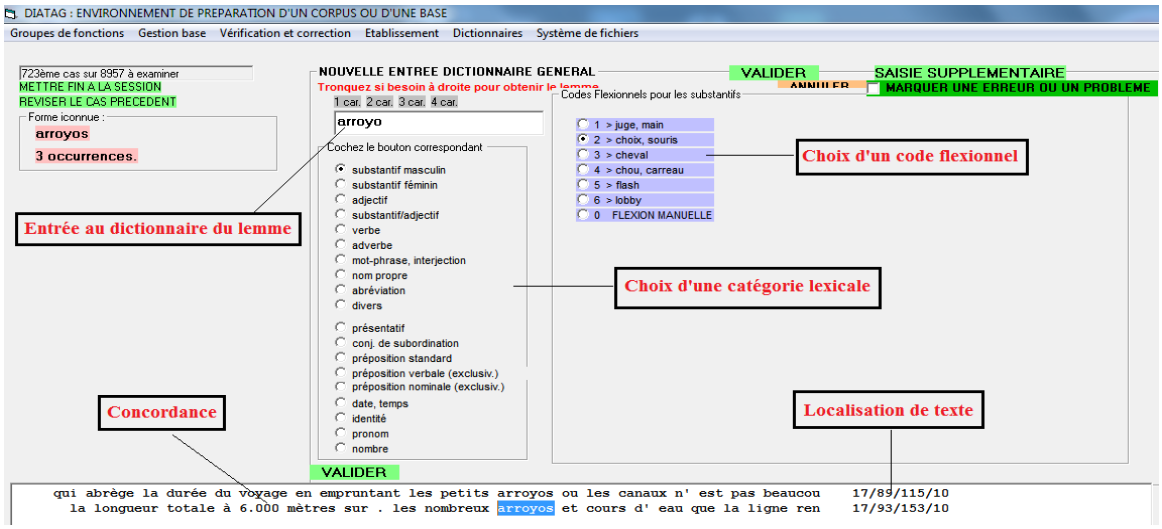

Figure 6 : Interface d'identification d'une forme non reconnue

Les formes marquées comme relevant de séquences composées sont saisies dans son intégralité. Elles vont dans les dictionnaires des mots composés ou des noms propres composés. Les formes composées que le logiciel n'a reconnues qu'en partie seront reconnues dans leur intégralité afin d'éviter de provoquer de nouvelles ambiguïtés.

DiaTag génère des entrées avant de les intégrer aux dictionnaires dans des fichiers nommés MANQUANT, éditables en mode texte. Il dispose d'un fléchisseur permettant de les annoter avant de les importer au dictionnaire. Chaque nouvelle entrée au dictionnaire est annotée d'une valeur flexionnelle[91]. Ainsi, les formes intégrées auront un caractère universel. Chaque intégration au dictionnaire va l'aider à s'agrandir.

En ce qui concerne les formes non reconnues provenant d'une mauvaise reconnaissance des caractères, DiaTag corrige déjà une partie lors de la correction immédiate en appliquant son

[90] Le dictionnaire du corpus est utilisé au même titre que celui du système. Nous l'utilisons lors d'une acquisition d'une forme temporaire ou afin d'éviter de fausser le dictionnaire du système

[91] Le lemme inclut l'indication de catégorie même si celle-ci n'est pas ambiguë ; le format est du type NOTER,.V:F ; NOTES,1.SF:P/1R.V:2K ; NOTEZ,1R.V:5N ; NOTÉE,2ER.V:YFS ; NOTÉS,2ER.V:YMP, NÉPAL,.N

rescor sur le corpus pour repérer les formes non reconnues récurrentes rencontrées dans les différents corpus. Ces formes sont reconnues par leur récurrence et par la reconnaissance de la proximité des traits principaux d'un caractère que l'océrisation a pu produire pendant l'identification des lettres, tels : *parli* pour *parti*, *pohce* pour *police*, *paroe* pour *parce*, etc. Ainsi, le dossier *rescor* est construit du fur à mesure pendant l'application du logiciel.

D'autre part, elle recense des formes non reconnues de tous les dictionnaires du logiciel en les indexant dans le dossier MANQUANT du logiciel où il y a entre autre un fichier des concordances de chaque forme recensée.

A chaque opération, DiaTag conserve le dossier avant modification et crée un nouveau dossier où est enregistré le dossier déjà modifié. Dans ce dernier dossier, les formes repérées auront un nouveau connecteur de forme fx : *<M f=boursiers fx=bouisiers >bouisiers*. Ce connecteur a un caractère temporaire. Une fois les formes validées, ce connecteur disparaitra.

A chaque opération, nous devrons fixer un seuil *n* qui fait référence à un nombre d'occurrences et qui permettra de prendre une décision quant au traitement des formes non reconnues. C'est-à-dire, que seules les formes non reconnues possédant plus de *n* (seuil) occurrences seront traitées. Le fait de fixer un seuil des occurrences à corriger limite le caractère aléatoire des erreurs recensées et améliore l'efficacité du processus. L'intérêt de travailler sur une grande base porte sur la gamme de fréquences du vocabulaire. Plus la base est grande, moins il y a d'hapax (en proportion). Plus le seuil des occurrences d'un item recensé est élevé, plus le risque d'erreur est faible. En outre plus les formes non reconnues sont corrigées ou réintégrées dans le dictionnaire du système, plus ce dernier sera complet.

4.2 Recensement et validation des fautes corrigibles

Les formes corrigées lors de l'application du dictionnaire sont recensées et générées dans un fichier éditable dans un tableur Excel pour faciliter la lecture de cette liste. Notre travail est de valider les formes proposées. L'idée est simple : le fichier est lu dans un tableur Excel où les propositions retenues sont cochées, ou bien de nouvelles formes sont proposées. Le fait de cocher des bonnes propositions permet de les trier plus facilement à la fin de la validation.

Assence	5298	absence	essence
Soldals	5242	soldats	x
Educition	5226	éducation	x
Ixtrême	5194	extrême	x
Nouvea	5149	nouvel	
Nouveu	5149	nouvel	
Souvel	5149	nouvel	
Ripports	5125	rapports	x
Critiqlie	5102	critique	x

Nombre d'occurrences d'un item dans le corpus de référence

Figure 7 : Exemple du tableur de validation des formes corrigées

Les candidats à valider doivent être des items complets pour s'assurer qu'ils ne soient pas coupés. Par exemple : *contrair* ne peut pas être validé en tant que contraire, car il peut être issu du mot *contrairement*. Ou bien *molion* ne peut pas être validé comme *émotion*,il peut

s'agir du mot *promotion* etc. Dans certains cas, les caractères manquants n'empêchent pas la validation des items, comme la forme *dochine*.

Dans le cadre de notre corpus, le travail de validation est réalisé sur l'ensemble des formes non reconnues possédant 2 occurrences ou plus. A la fin de l'opération, les propositions retenues et les nouvelles formes proposées sont importées dans le *rescor*. Ainsi, plus le logiciel est propagé, plus son *rescor* est riche. Par conséquent, il couvre plus de formes non reconnues. A la sortie de cette opération, nous appliquons de nouveau la correction automatique sur le corpus.

4.3 Passage corrompu

Vu le nombre d'hapax restant à corriger (plus de 7% des occurrences du corpus), nous pouvons imaginer qu'il y a beaucoup de passages corrompus où la correction automatique, et même une correction via l'interface de DiaTag est impossible. Cela pour deux raisons : une ressource en temps beaucoup trop importante et une fiabilité non assurée. Ainsi, pour ces passages, DiaTag nous propose de les ressaisir manuellement.

Sur l'interface de l'opérateur, nous avons côte à côte l'article en PDF original (d'où l'intérêt de découper le journal en article) et une fenêtre de communication où nous saisissons le texte. Ce travail n'est pas très scientifique, mais en attendant de trouver une solution pour traiter les archives qui atteignent une importante dégradation, nous préférons opter pour la méthode la plus sûre et donc améliorer la fiabilité des interrogations ultérieures sur les données textuelles.

5. Bilan et perspectives

5.1 Le mal des outils informatique performants

Les difficultés rencontrées lors de la constitution du corpus *La Tribune Indochinoise* nous révèle les lacunes en outils performants pour pouvoir traiter ce genre de document.

Nous avons pourtant constitué ce corpus avec un des logiciels les plus performants existant à ce jour dans ce domaine. Mais il est développé comme la plupart des programmes actuels pour une lecture de données textuelles fixées sur un document de bonne qualité et dont le vocabulaire est plus restreint. Face aux documents anciens dont les défauts sont multiples (les espaces fluctuants qui provoquent souvent une lecture ambiguë, des bruits provenant de la dégradation des originaux ou encore la technologie d'imprimerie de l'époque …), et surtout un vocabulaire riche. Il a donc réalisé un taux de reconnaissance insuffisant.

Nous espérons que dans le futur, nous obtiendrons un logiciel qui sera capable de reconnaitre tous les caractères et surtout armé d'un dictionnaire encore plus riche en vocabulaire.

5.2 Le corpus La Tribune Indochinoise: l'innovation ou la tradition ?

Aurions-nous eu une possibilité de constituer autrement ce corpus ? La réponse est non pour deux raisons. Premièrement, nous ne sommes pas une bibliothèque qui cherche à construire une base de données numérique. Nous sommes un laboratoire sémiotique, linguistique, didactique et informatique. Nous cherchons donc à constituer des données textuelles à des fins d'analyse textuelle des discours. Il est important que nos données textuelles soient présentées

en mode plein-texte et avec une bonne fiabilité afin de ne pas biaiser le traitement statistique ultérieur. Deuxièmement, il s'agit de presse ancienne scannée à partir de microfilms. Elle accumule un grand nombre de défauts rendant impossible un traitement efficace par logiciel. Nous risquions de passer beaucoup de temps et de moyen pour effectuer ce genre de corpus. Or le temps est précieux dans les recherches et les moyens ne sont pas toujours disponibles. Ainsi avec le résultat obtenu pour ce corpus, nous pourrions dire que notre méthode employée dans ce travail a été innovante.

Pour conclure, constituer un corpus linguistique à partir de documents anciens n'est pas un nouveau défi. Mais chaque document présente des difficultés différentes. Donc chaque corpus a son propre mode de constitution. Ainsi, chaque constitution de ce type de corpus restera un nouveau défi.

Bibliographie

CHAUMIER, Jacques (2006). *Document et numérisation : enjeux technique, économiques, culturels et sociaux*, Paris, ADBS, 119p.

JACQUESSON, Alain ; RIVIER Alexis (2005). *Bibliothèques et documents numériques : Concepts, composantes, techniques et enjeux*, Paris, Editions du Cercle de la Librairie, 573p.

LETHIER Virginie (2009). *Exploration textuelle du discours d'un quotidien régional au carrefour du XIXe et du XXe siècles : Le Petit Comtois (1883-1903).*

POLIS, St ; STASSE, B. (2009). *Pour une nouvelle philologie numérique : réflexions sur la relation texte(s)-document(s),* [http://popups.ulg.ac.be/MethIS/docannexe.php?id=291].

VIPREY, Jean-Marie (2005). « *Philologie numérique et herméneutique intégrative* » in *Sciences du discours en dialogue : Textualité & comparaison,* dir. Jean-Michel Adam, Claude Calame, Ute Heidmann, SlatkineInternet et base de données.

http://laseldi.univ-fcomte.fr/document/viprey/page_JMV.htm