PALM-Méditext

Une plateforme internet pour l'analyse linguistique de textes médiévaux (PALM)
Un corpus de textes politiques d'origine anglaise et française de la fin du Moyen Âge (Méditext)

J.-P. Genet, C. Fletcher, A. Mairey, N. Kanaoka, C. Morgan, L. Albiero, M. Aouini

PALM est un système informatique en ligne qui permet l'application de logiciels d'analyse linguistique aux textes médiévaux. La version initiale a été conçue pour traiter des textes politiques d'origine française et anglaise, en latin médiéval, moyen français ou moyen anglais, couvrant la période entre le XII^{e} et le début du XVI^{e} siècle, tout en laissant ouverte la possibilité d'ajouter par la suite d'autres thèmes, d'autres langues et d'autres régions. Avant PALM, il était impossible, ou au moins humainement très exigeant, de préparer un grand corpus de textes de cette période pour l'analyse par des logiciels d'analyse textuelle, à cause de l'absence d'orthographe standardisée et également, dans le cas du moyen anglais, de la nature toujours changeante de sa grammaire et de sa syntaxe.

PALM fait partie d'un programme beaucoup plus vaste dirigé par J.-P. Genet et intitulé *Signs and States : Les vecteurs de l'idéel.* La mission de ce projet ERC est d'examiner les multiples manières par lesquelles le développement et l'application de nouvelles formes de gouvernance et de pouvoir étatique à la fin du Moyen Âge ont transformé les sociétés européennes, non seulement socialement et économiquement, mais également culturellement. Ce projet se concentre donc sur le pouvoir symbolique.

Les historiens ont souvent souligné le rôle transformateur de deux éléments centraux dans la formation de « l'État moderne » : le développement de la justice centralisée et de l'impôt public. Un peu partout en Europe, l'autorité caractéristique d'un roi est de plus en plus monopolisée par une seule autorité dans des territoires qui sont, d'ailleurs, de plus en plus précisément définis. Elle s'opère souvent au détriment des coutumes et des privilèges locaux, notamment par le biais de formules issues du droit romain qui permettent d'accroître le droit de ressort du prince et de justifier sa législation. Par exemple, l'argument de la nécessité évidente et l'élaboration de l'idée de représentation ont permis aux rois et aux pouvoirs princiers de lever de nouveaux impôts et de financer des armées d'une taille sans précédent, soit en faisant passer la justification de leurs actions par des assemblées représentatives, soit en appelant au bien commun. Toutefois, dans le même temps, des valeurs liées à la genèse de l'État coexistaient avec des systèmes de valeurs plus anciens, parfois en contradiction, parfois créant des phénomènes entièrement nouveaux.

De tels thèmes ont été analysés par les historiens depuis plusieurs décennies. Plus récemment, ils ont commencé à se concentrer plus précisément : sur les idées politiques populaires ; sur la « culture politique » ; sur « l'espace public » ; sur la politique envisagée moins comme la haute politique au niveau national, que comme un type d'interaction discernable à plusieurs niveaux sociaux ; et sur la nature de « l'État » à la fin du Moyen Âge. Ce dernier thème, en particulier, a été examiné dans les années 1980 et 1990 lors de plusieurs colloques et publications dirigés par J.-P. Genet. Les participants ont considéré dans un contexte comparatif

surtout, mais pas exclusivement, le développement de l'impôt, de la justice et de la « société politique » – l'ensemble toujours en mutation de ceux qui ont joué un rôle important dans la vie politique. Le projet ERC *Signs and States : Les vecteurs de l'idéel* nous a permis de franchir une nouvelle étape. Ce projet vise moins l'étude du développement des phénomènes étatiques en tant que tel que celle de leur impact sur les mentalités et leur interaction avec la culture telle qu'elle existe déjà. Le but est d'appréhender les idées reçues et les valeurs conscientes des populations qui ont été touchées par le développement de l'État et qui ont participé à son élaboration, en même temps que la nature des innovations venues « d'en haut ».

Une approche possible à ces problèmes est l'étude de la terminologie. Son histoire occupe aujourd'hui une place acceptée au sein de la recherche historique, notamment dans les travaux de « l'École de Cambridge », souvent associé à P. Laslett, Q. Skinner et J.G.A. Pocock, et dans la *Begriffsgeschichte*, suivant l'exemple de R. Koselleck et H. Gumbrecht. Toutefois, jusqu'à assez récemment, ces recherches ne pouvaient se concentrer que sur un seul auteur ou un groupe d'auteurs, ou sur un seul terme ou un groupe de termes. Dans le programme actuel, bien que les textes des auteurs traditionnellement étudiés sont toujours étudiés, l'objet principal de la recherche est plus ambitieux : la langue elle-même comme un système de signes qui est un des vecteurs par lesquels (suivant M. Godelier) « l'idéel » d'une société donnée est projetée. Il était impossible d'appliquer les méthodes statistiques qui permettent l'analyse globale du vocabulaire d'un grand corpus avant l'émergence très récente de ressources informatiques puissantes et à la portée de tous. Dans le cas des textes médiévaux, l'absence d'orthographe standardisée a rendu difficile même des recherches sémantiques simples, telle que l'observation de l'utilisation d'un terme à travers un grand corpus de textes.

D'où l'intérêt de PALM : elle rend possible l'exploration et surtout l'analyse sémantique et statistique des textes politiques de la fin du Moyen Âge.

Les documents suivants présentent le travail de l'équipe de ce projet ERC *Signs and States : Les vecteurs de l'idéel.*

1. Une introduction à PALM
2. Une liste des textes dans la « Librairie » de PALM (Méditext)
3. Une présentation technique de PALM
4. Contacts internationaux de l'équipe
5. Bibliographie de l'équipe
6. Présentation et minutes de la *Journée d'Etudes 'Méditext'*, 18-19 mai, 2011

1. Une Introduction à PALM

Qu'est-ce que c'est que PALM ?

PALM est avant tout une *plateforme* d'analyse linguistique médiévale, c'est-à-dire un outil qui permet de réunir les ressources informatiques existantes pour le traitement de textes médiévaux et de développer de nouvelles ressources linguistiques. En particulier, il est important de souligner que l'objectif de PALM n'est pas de fournir des analyses statistiques avancées. Sa fonction est de mettre les textes médiévaux dans une forme qui peut ensuite être traitée par des logiciels existants pour l'analyse linguistique de textes de toute origine. Nous ne voulons pas « réinventer le fil à couper le beurre ». Pour le moment, PALM peut exporter des textes vers des logiciels développés en France, tels que Lexico 3, Hyperbase et TXM, mais sa structure est conçue pour permettre plus tard l'export vers des outils existants ou à venir.

De manière générale, il est nécessaire de passer par trois étapes avant d'entreprendre l'analyse informatisée d'un corpus de textes médiévaux :

1. la composition du corpus ;
2. la « lemmatisation » – la plus difficile des étapes ;
3. la mise-en-forme du corpus pour traitement par un logiciel spécifique.

PALM vise à faciliter ces trois tâches, et à les automatiser autant que possible. PALM est une plateforme en ligne, accessible par un login et un mot de passe, disponible sur demande. L'adresse web est : http://lamop-vs3.univ-paris1.fr/PALM.

Comment utiliser PALM ?

1. La composition d'un corpus de travail

En premier lieu, PALM permet de constituer un corpus de travail, soit à partir de textes que l'utilisateur aura lui-même fourni, soit à partir de la « Librairie » de la plateforme [fig. 1]. C'est la constitution de la « librairie » de PALM qui a occupé l'essentiel de la première année de notre programme, d'octobre 2010 à octobre 2011 approximativement (après une phase de recrutement de six mois). Les autres réalisations de cette période ont été l'adaptation des ressources existantes à Paris I (notamment textuelles) et la recherche de nouveaux textes à inclure.

Les origines de cette librairie remontent à « Méditext » : un corpus de textes d'origine anglaise et française rassemblés depuis les années 1980 sous la direction de J.-P. Genet et Cl. Gauvard.

Ce corpus rassemble essentiellement des textes « politiques », c'est-à-dire :

- soit des textes ayant trait à des événements politiques identifiés ;
- soit des textes consacrés de manière générale au bon et au mauvais gouvernement.

Pour le moment, la « Librairie » de PALM regroupe des textes d'origine anglaise (en anglais, français et latin) et d'origine française (en français et en latin). Toutefois, nous souhaitons, à l'avenir, ajouter des textes en provenance d'autres pays européens, et la plateforme est conçue pour permettre cette expansion.

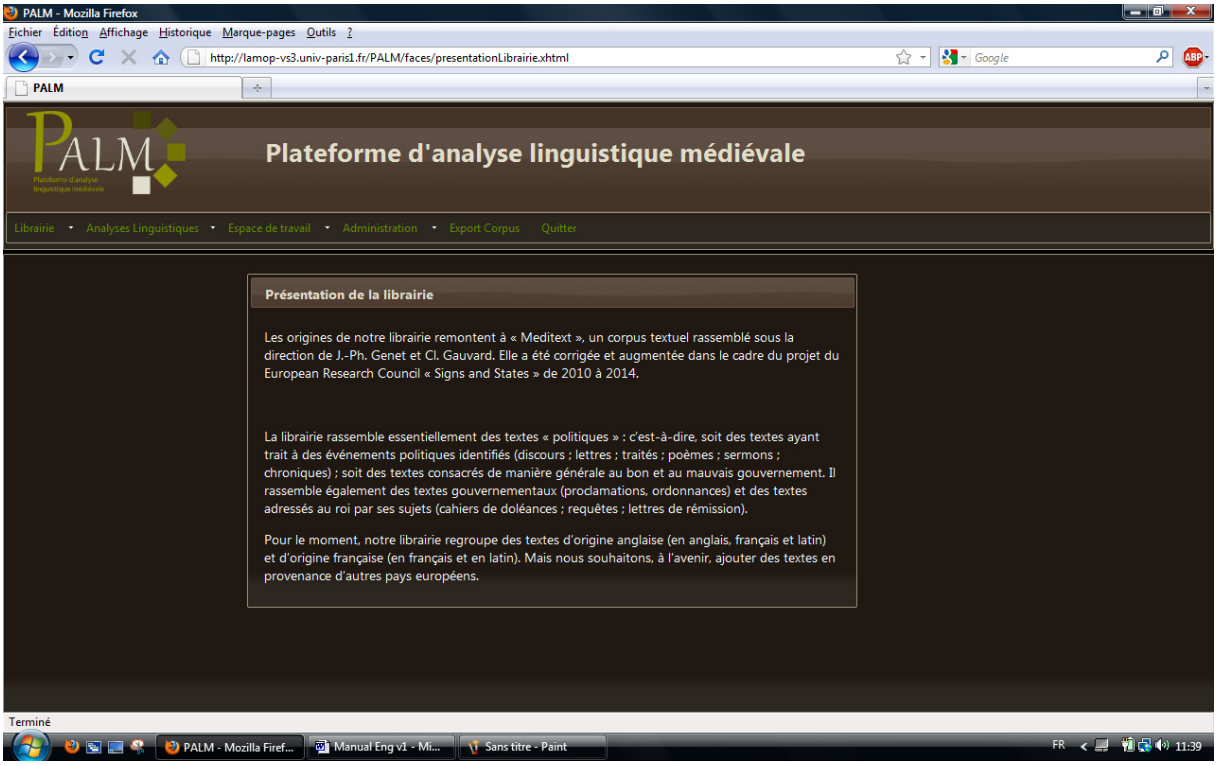

Fig 1: Page d'accueil de PALM: Présentation de la « Librairie » intégrée

Méditext et la Librairie intégrée de PALM

Pour une présentation détaillée des textes de la Librairie de PALM, vous pouvez consulter la liste de textes attachée en annexe. Quelques remarques, toutefois, sur la nature des textes :

Textes d'origine française

Les textes d'origine française sont surtout en français et, dans une moindre mesure, en latin. Il s'agit d'abord d'un corpus de traités politiques : le *Quadrilogue Invectif* d'Alain Chartier, par exemple, ou le *Songe du vieil pèlerin* de Philippe de Mézières. Ensuite, nous avons deux corpus saisis sous la direction de Claude Gauvard : un corpus d'*ordonnances* de Charles VI ; et un corpus de lettres de rémission – documents qui accordent un pardon royal et qui, ce faisant, incorporent les suppliques qui sont à leurs origines, souvent de manière très détaillée. Il s'agit donc surtout de textes de la fin du XIV^{e} et du début du XV^{e} siècle.

Au début du projet, ces textes étaient préservés dans une grande variété de formats, caractéristiques de l'époque de leur numérisation. Les plus anciens ont été saisis sur des cartes perforées, qui ont ensuite été scannées. Cela peut paraître une perte de temps si l'on ne prend pas en compte les limites des logiciels de reconnaissance optique de caractères (OCR). Dans l'état actuel de la technologie, ces logiciels fonctionnent très mal pour les langues médiévales (ils n'ont pas de dictionnaires de contrôle adaptés) ainsi que pour les éditions anciennes, comme par exemple celles du XIX^{e} siècle (ils peinent à reconnaitre les polices et le formatage de l'époque).

Le travail de mise en forme des textes d'origine française dans Méditext a été considérable. Il a représenté plus d'une année de travail pour notre linguiste spécialisée en moyen français, Naomi Kanaoka (d'octobre 2010 jusqu'à la fin de l'année 2011).

Récemment, Mme Kanaoka a commencé à accroître ce corpus, en commençant par l'ajout d'un ensemble de traités politiques et de poèmes politiques en français. Elle a notamment enrichi la librairie de PALM en corrigeant des versions électroniques externes de la version française du *De Regimine Principum* de Gilles de Rome, un traité politique très diffusé à la fin du Moyen Âge. Nous avons également pu inclure un grand corpus des sermons français de Jean Gerson. Nous avons l'intention d'augmenter ce corpus en ajoutant une sélection de textes historiques, issue notamment des *Grandes chroniques* de France et de la vieille édition des œuvres de Froissart.

Textes d'origine anglaise

Il faut souligner que les textes d'origine anglaise sont en trois langues différentes : le français, le moyen anglais et le latin. De la Conquête Normande jusqu'à la fin du XII^{e} siècle, le latin domine la pratique de l'écrit en Angleterre. Dès le milieu du XII^{e} siècle, le français commence à acquérir le statut d'une langue littéraire, et, dès la fin du XIII^{e} siècle, de la langue dans laquelle il est approprié de retranscrire les procès, d'écrire des lettres, des proclamations publiques et des requêtes. Notons l'ironie : ces développements prennent place exactement au moment où la langue anglaise écrite émerge, d'abord comme moyen d'instruction religieuse et de divertissement populaire, et, de plus en plus, dès la fin du XIV^{e} siècle, pour l'archivage (minutes, comptes, récits historiques...), pour des œuvres littéraires élaborées (Chaucer, Gower, *Piers Plowman*...), et pour des lettres, des requêtes et des traités politiques. Pourtant, même à la fin du Moyen Âge, le français, le moyen anglais et le latin, sont encore employés simultanément, parfois dans le même document.

Les textes du corpus Méditext d'origine anglaise se divisent en trois groupes principaux :

- un corpus de discours, de traités, de proclamations royales et de lettres associés à des moments politiques spécifiques entre le XIII^{e} et le XVI^{e} siècle, en trois langues, issus des archives de la monarchie anglaise ou des collections privées, notamment des textes ayant trait au Parlement rassemblés par J.-P. Genet.
- un corpus de poésie politique, encore une fois lié à des moments politiques, recueilli par J.-P. Genet et par A. Mairey.
- des sermons et des traités politiques en anglais, du XIV^{e} jusqu'au XVI^{e} siècle.

Depuis octobre 2010, ce corpus a été augmenté, d'abord par notre première linguiste pour le moyen anglais, Rachel Moss, et depuis novembre 2011, par Chloë Morgan. Nous avons rajouté des textes politiques en anglais, notamment la traduction anglaise du *Regimine Principum* de Giles de Rome ; des textes politiques du XV^{e} siècle recueillis dans le livre de John Vale ; un corpus de textes historiques en anglais, notamment le *Brut*, et la traduction de John Trevisa du *Polychronicon* de Ranulph Higden.

En même temps, nous avons également ajouté des textes d'origine anglaise mais en langue française ou en latin, par exemple : de la poésie politique d'origine anglaise en latin ou en français des XIII^{e} et XIV^{e} siècles ; et des textes d'ordre politique : des proclamations royales, des lettres ouvertes telles que le *Libellus famosus* d'Adam d'Orleton, et des ordonnances et des traités politiques.

Quelques mots sur les textes en latin

À l'origine, notre corpus latin de textes politiques de la fin du Moyen Âge était bien plus restreint que ceux des deux langues vernaculaires. Les textes dont nous disposions étaient surtout des proclamations royales d'origine anglaise, des lettres, des discours et des sermons issus des archives du Parlement anglais. Nous avions également la chronique du religieux de Saint Denis. En outre, nous avons eu quelques problèmes de recrutement dans le cas du latin. Notre première latiniste, une étudiante en thèse, n'a commencé qu'en janvier 2011 et n'est restée dans notre équipe qu'un an.

Ainsi, mis à part le travail de correction du corpus Méditext et sa préparation pour PALM, le travail de notre première latiniste a consisté à saisir ou nettoyer de nouveaux textes, et en particulier des textes en latin très diffusés au niveau européen. D'un côté, elle a corrigé des traités politiques, saisis extérieurement, tels que le *Policraticus* de Jean de Salisbury. De l'autre, elle a ajouté des poèmes politiques en latin et quelques textes polémiques liés, par exemple, aux crises de la première moitié du XIV^{e} siècle en Angleterre. La numérisation de la version latine du *De Regimine Principum* de Gilles de Rome a été financé grâce à SAS, et nous avons reçu le *De Regno* de Thomas d'Aquin de l'université de Nancy. Nous avons commencé leur correction.

Au mois de janvier 2012, nous avons recruté Laura Albiero. En plus de la création et de l'adaptation de ressources linguistiques, elle a commencé à préparer pour PALM des textes historiques, notamment la version latine du *Polychronicon* de Ralph Higden, et des traités politiques liés à la controverse entre Philippe IV et Boniface VIII.

Comment créer un corpus avec PALM ?

La première grande tâche de notre ingénieur en informatique linguistique, Mourad Aouini, a été de construire un environnement pour la consultation et la manipulation de la « Librairie » constituée à partir de Méditext, en moyen anglais, moyen français et latin. Les aspects techniques de cette tâche sont décrits dans l'annexe appropriée.

Une fois passée l'étape du login et du mot de passe, l'utilisateur de PALM peut parcourir les textes dans la « Librairie » [fig. 2]. Il peut faire des recherches par mot clé dans chaque champ, ou en mettant la liste de textes en ordre alphabétique dans une sous-catégorie, par exemple par langue, ou par pays d'origine.

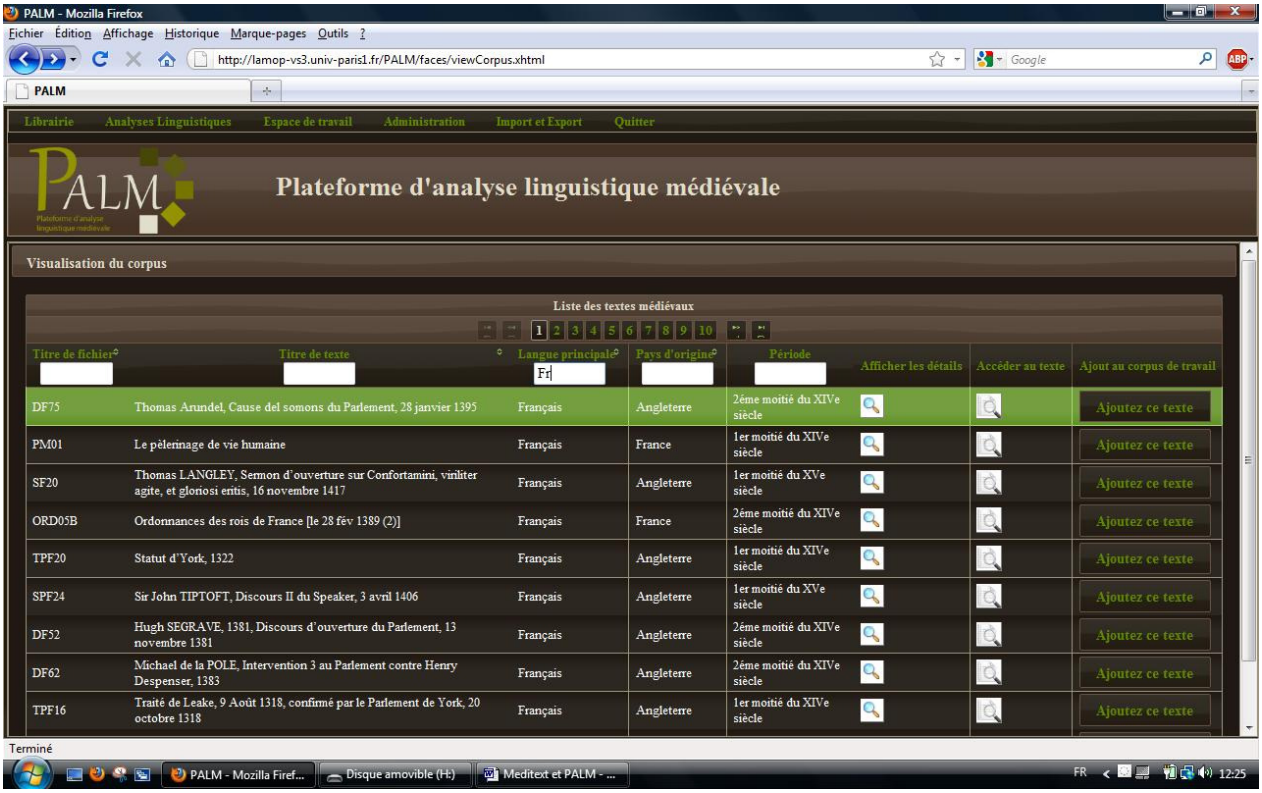

Fig. 2: La « Librairie » organisée par langue principale : français

Il est également possible de consulter une brève description de chaque texte. Ici, par exemple, les détails de trois ballades politiques de Jean Creton, écrivain de la fin du XIV^{e} et du début du XV^{e} siècle [fig. 3].

Fig. 3: Détails: Trois poèmes de Jean Creton

Et on peut consulter le texte [fig. 4].

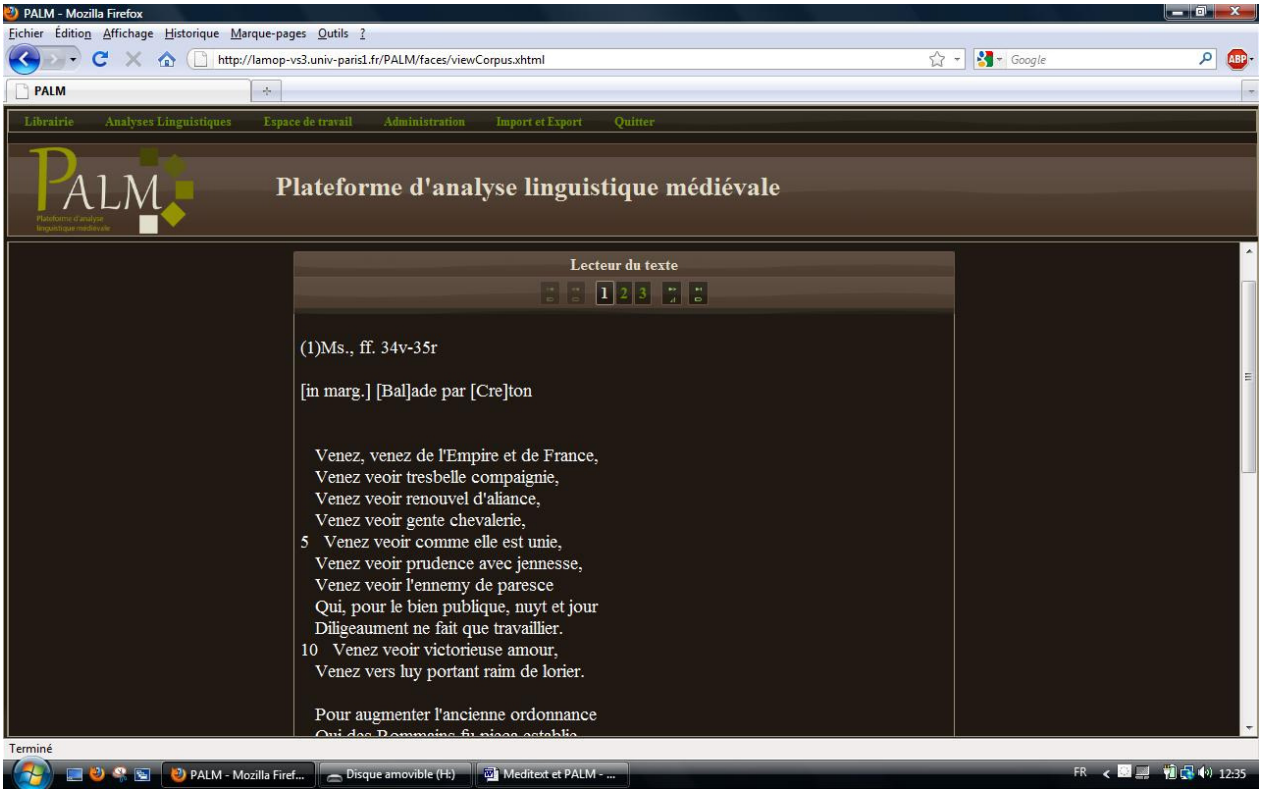

Fig 4: La « Librairie » : consultation du texte

À chacune de ces étapes, l'utilisateur peut décider de sélectionner un texte pour le mettre dans son propre corpus – son « Corpus de travail » – pour ensuite l'utiliser lors des étapes suivantes (lemmatisation, exportation). Il est possible de constituer un corpus uniquement sur la base des textes qui sont déjà dans la Librairie. Mais l'utilisateur peut également ajouter ses propres textes à son « Corpus de travail ». Pour le moment, cette étape se fait par le biais d'un formulaire, accessible dans le menu Librairie : « Ajouter un texte ».

Imaginons que l'utilisateur désire analyser un échantillon de requêtes soumis au roi par les *Commons* lors du Parlement anglais en janvier 1437 [fig . 5]. Il doit d'abord entrer le descriptif du texte : un « nom de fichier » court ; un titre plus long et descriptif ; les détails de l'édition du texte ; la date, etc.

Il peut également spécifier son niveau d'accès : tous les utilisateurs ; accès restreint ; ou administrateur seulement – par exemple dans le cas d'une édition en préparation.
Dans la version présentée dans la fig. 5, il fallait ajouter le texte en le collant dans la fenêtre « page », mais nous avons maintenant mis en place la possibilité d'importer dans PALM un fichier texte ou Word.

Si l'utilisateur le souhaite, il peut marquer la pagination avec des balises de style <p=1>.

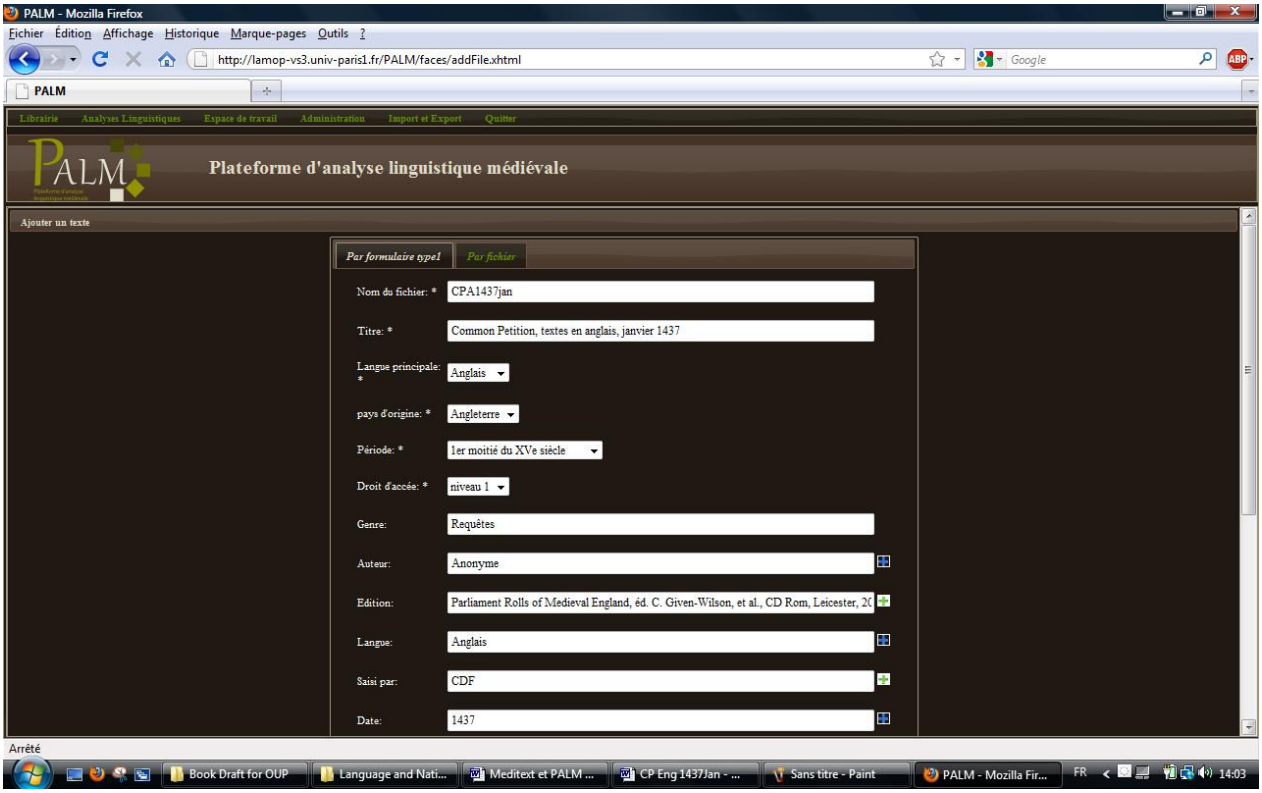

Fig. 5: Comment ajouter un texte au « Corpus du travail » de l'utilisateur.

Si l'utilisateur clique sur « Envoyer le texte », le texte est téléchargé dans son corpus de travail. Il peut le consulter dans la liste des ses textes. Il peut regarder les détails ou le texte lui-même.

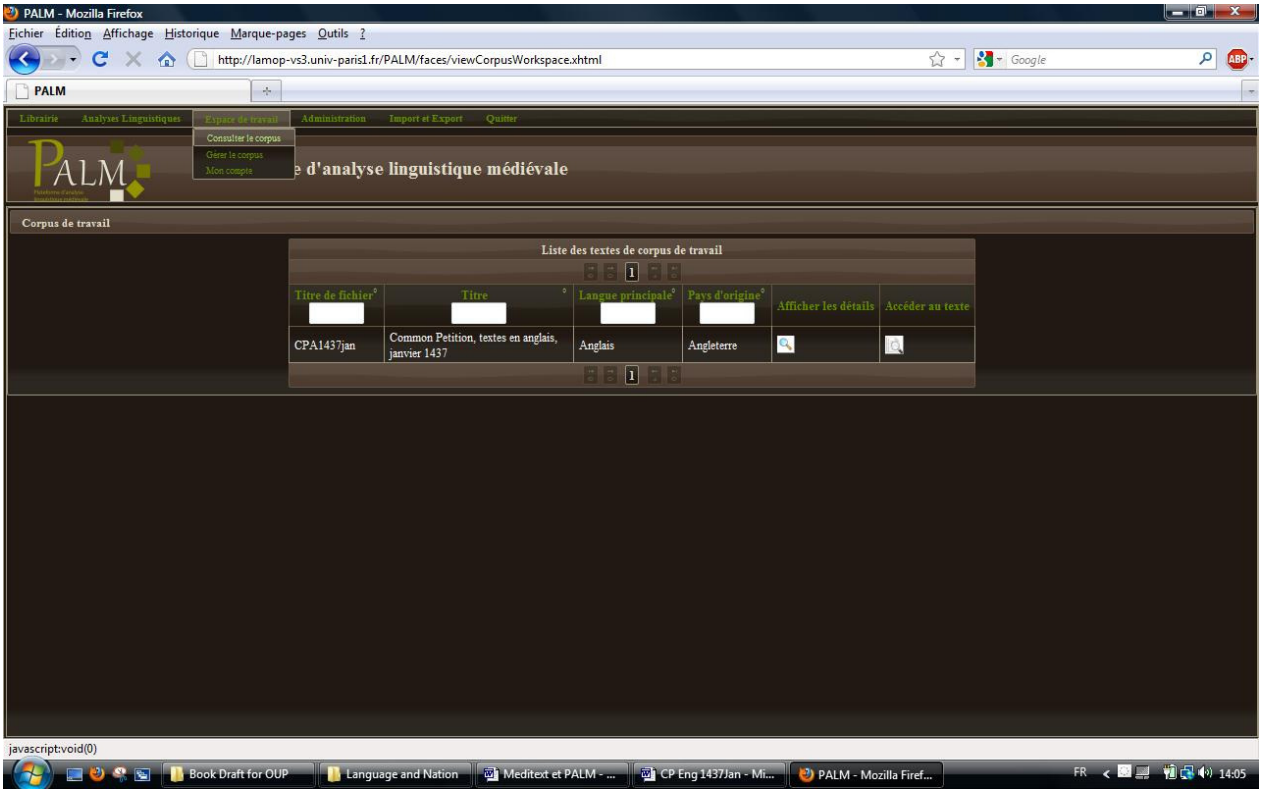

Fig 6: Un corpus de travail d'un seul texte.

Imaginons que l'utilisateur souhaite comparer ce discours avec d'autres textes de la même époque, dans le sillage du traité franco-bourguignon conclu à Arras en 1435 et du siège de Calais de 1436. Il revient à la librairie, il sélectionne ces textes, et il clique « Ajoutez ce texte » à chaque fois. Ses choix apparaissent dans le corpus de travail, et il peut passer à l'étape suivante.

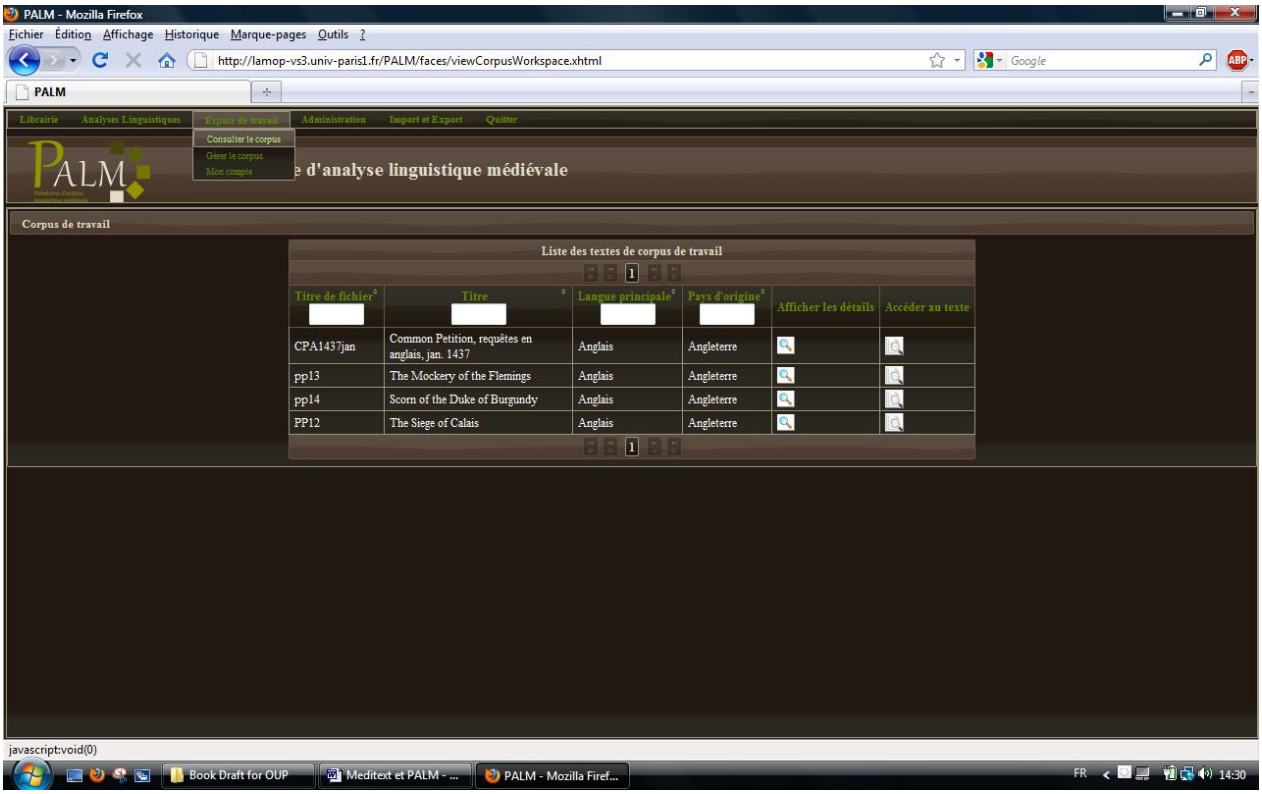

Fig. 7: Un petit corpus prêt pour la lemmatisation.

À partir du menu « Espace de travail », l'utilisateur peut modifier les détails de ces textes, ou les enlever de son corpus. Le menu « Administration » le permet de changer son mot de passe et ses détails personnels.

Ainsi peut-on voir que PALM rend déjà plus aisée la composition d'un corpus et donne accès à une « Librairie » de textes au sein de la plateforme. C'est la première étape. Pourtant, elle est loin d'être la plus difficile. Nous sommes actuellement en train de passer à la prochaine étape : celle de la lemmatisation.

2. Lemmatisation

En effet, pour utiliser les outils informatiques d'analyse des textes médiévaux, il est indispensable, pour obtenir de bons résultats statistiques et même pour effectuer des recherches efficaces, de les « lemmatiser ». Qu'est-ce que c'est que la lemmatisation ? Il s'agit là d'annoter chaque forme dans un texte avec son « lemme », c'est-à-dire avec la forme canonique de ce mot telle qu'on la trouverait dans un dictionnaire de cette langue.

Les bienfaits de la lemmatisation varient selon la langue et selon la période historique du texte considéré. Dans le cas du français moderne, par exemple, son principal avantage est de regrouper toutes les formes fléchies des noms, des verbes et des adjectifs. Ainsi, par exemple, pour *vous aimez*, *nous aimons* et *tu as aimé*, le verbe dans chaque cas est annoté par l'infinitif *aimer*. Le

nom dans *deux chevaux* ou *un cheval* est relié dans les deux cas à la forme du dictionnaire : *cheval*. Dans une langue fléchie, la lemmatisation permet donc de traiter en même temps, par exemple, toutes les formes d'un verbe ou d'un nom.

Dans le cas des langues médiévales, et surtout pour les langues vernaculaires, la lemmatisation est encore plus nécessaire, dans la mesure où il n'existe pas d'orthographe standardisée. Ainsi, dans le cas du moyen français, si on consulte le *Dictionnaire du Moyen Français* en ligne, qui est consacré à la langue française des XIII^e^ et XV^e^ siècles, on trouve plusieurs orthographes différentes, même pour les mots les plus simples. Par exemple, dans le cas du verbe *aimer*, le *DMF* donne trois variantes – *amer, aimer* et *aymer* – ce qui multiplie par trois les orthographes possibles dans les formes fléchies de ce verbe.

Les mots plus longs montrent encore plus de variation. Ainsi, pour *léopard* le DMF donne *liepart*, *lieppart*, *leoppart*, *lupar* et *lupart* au singulier et *liepars*, *lieppars*, *limpars* et *luppars* au pluriel. Même pour le nom ou titre *seigneur*, pourtant très fréquent dans nos textes, le DMF liste neuf orthographes possibles au singulier et cinq au pluriel : *sainieur, segneur, seigneur, seignour, seigny, seingneur, siegneur, signeur* et *singneur* ; *saingnours, scenours, segneurs, seigneurs* et *seingners*.

La situation pour le moyen anglais est encore pire. Ainsi, si nous prenons un seul texte, le *Ayenbite of Inwyt*, un texte religieux du deuxième quart du XIV^e^ siècle, et si nous cherchons toutes les formes observées dans les textes annotés du *Linguistic Atlas of Early Middle English* (LAEME), par exemple pour le mot *kued*, qui donne en anglais moderne *wicked* (c'est-à-dire, méchant, comme pour la méchante sorcière), on trouve 10 formes : *kuead, kueade, kvead, quead, kued, queade, kuade, kuede, kueades, kuedes*. Ou, dans le même texte, pour *saint*, 16 forms: *saynt, zaynte, saint, sayn, zayte, seynt, zainte, sanyn, sanyt, saynte, seint, zainte, zaynte, zayn, sayn* et *sauynt*.

En outre, si on prend l'exemple du pronom réflexif, « *himself* », qui peut prendre plus simplement la forme « *him* », et si on le recherche dans tous les textes annotés par l'équipe du LAEME, nous trouvons quelques 90 formes pour ce lemme seul : *him-selue, him-seluen, him-self, himself, him-sulf, him-selu, him-seolf, himseolf, him-seoluen, hym-self, him-zelf, him-zelue, him-seolue, himselue, himseluin, him-silf, him-sulue, him-seolf, him-solf, him-solue, himm-sellf, himm-sellfenn, himmsellfenn, himmsellf, himm-sellf, ym-self, himseluen, he-sulf, hymself, him-selen, he-seolf, him ysself, him-selfen, im-self, him-seoluen, hine-selue, him-selfi, him-sulfne, him-suelf, him-sulne, hine-seolfe, hine-seolfne, hine-seolf, hine-solf, hm-solf, him-seelf, hine-sulfne, hine-seulfne, hine-sulue, him-suluen, himseoluen, him-seluein, hym-selue, him-seluen, him-sulfen, him-selfe, himselfen, hine-sulne, hine-suluen, him-sylfe, him-seluum, him-silfum, him-soluen, him-seolfne, him-seoluan, him-selua, him-silue, himzelue, hym-selwe, him-selwen, him-seoluen, him-seluan, him-seolfe, himsuluen, him-seolue, hire-seolue, him-sulf, him-suluen, himm-sellfenn, himseluen, hym-syfe, hine-silfne, him, himm, hym, im, hem, hine, hym, hyne, hine*.

Remplacer tous les mots d'un texte par leur lemme, ou même par une orthographe standardisée, constitue une tâche énorme. Même pour des recherches simples, il faut faire plusieurs recherches pour toutes les variantes orthographiques. Normalement, pour des analyses statistiques, l'historien doit, soit décider à l'avance quels mots l'intéressent, soit choisir les plus fréquents, soit se limiter à des textes courts, ce qui rend moins signifiants les résultats de l'analyse finale.

L'objectif principal de la plateforme PALM est donc à la fois de mettre à la disposition de l'utilisateur tous les outils électroniques existants pour faciliter la lemmatisation d'un texte, et de créer de nouveaux outils linguistiques pour que cette lemmatisation soit plus rapide et plus précise.

Les ressources linguistiques

À mi-chemin de ce projet, la mise en place de la lemmatisation représente le principal aspect de notre travail. Le problème principal est qu'il n'existe pas les mêmes ressources linguistiques informatisées pour le latin, le moyen français et surtout pour le moyen anglais, ce qui n'est pas le cas pour les langues modernes. Pour le moment, nous nous concentrons sur deux types de ressources linguistiques informatisées pour faciliter la lemmatisation. Ils seront présentés brièvement dans cette annexe, et de manière plus détaillée dans l'annexe technique.

a) Les dictionnaires électroniques

Pour les langues modernes, il existe de simples listes de formes, annotées par leur lemme et par leur catégorie grammaticale, appelées généralement « dictionnaires électroniques » ou, plus précisément, DELAFs – Dictionnaires ELectronique de Formes fléchies. En appliquant un dictionnaire de ce type à un texte en langue moderne, l'ordinateur peut suggérer plusieurs possibilités d'annotation pour chaque mot.

Mais l'application d'un DELAFs a ses limites. Une telle méthode traite un texte mot par mot, sans prendre en compte le contexte grammatical. Elle ne peut pas résoudre des cas ambigus qui ne posent pourtant aucun problème à un lecteur humain. Ainsi, par exemple, plusieurs formes peuvent-elles remonter, soit à un nom, soit à un verbe. Par exemple, en français moderne : « Il a marché sur la lune » ou « Il est allé au marché » ; ou, en anglais : « Don't step on my toe ! » ou « He swept the front step ». Un être humain parlant cette langue distingue sans problème entre un verbe et un nom qui ont la même forme, mais un dictionnaire de formes, qui traite un texte mot par mot et non phrase par phrase, ne peut pas le faire.

Il faut donc faire appel à une deuxième ressource informatisée : un « taggeur ».

b) Les « taggeurs »

Les « taggeurs » sont des logiciels qui fonctionnent à partir de l'intelligence artificielle. Ils essaient de résoudre les cas ambigus grâce à un entraînement préalable sur des textes qui ont déjà été annotés par des utilisateurs humains. Les taggeurs peuvent donc résoudre les ambiguïtés grammaticales en proposant la solution la plus probable.

Le problème majeur est qu'il n'existe pas de dictionnaire électronique pour le moyen anglais et que personne, à notre connaissance, n'a jamais entraîné un taggeur sur cette langue. La situation est encore plus sérieuse que pour les autres langues de notre corpus, dans la mesure où la variation orthographique en moyen anglais est très marquée. En outre, la grammaire du moyen anglais est encore très éloignée de la grammaire standardisée de l'anglais des XVII^{e} et XVIII^{e} siècles, par exemple. La création d'un dictionnaire électronique de formes fléchies est un travail de plusieurs décennies.

Le cas du moyen français est un peu moins ardu. L'équipe du *Dictionnaire du Moyen Français* à l'université de Nancy ont crée un logiciel (LGerM) qui suggère plusieurs solutions à un utilisateur lui présentant une forme inconnue. Néanmoins, on ne peut pas dire que ce logiciel lemmatise automatiquement. C'est l'utilisateur qui choisit finalement entre plusieurs suggestions. Il est cependant possible, grâce à l'équipe du DMF, de faire passer un corpus de textes par ce logiciel, d'en extraire les bonnes réponses, et de constituer ainsi la base d'un DELAF. D'autres équipes ont entrainé des taggeurs sur l'ancien français, mais la langue de cette époque est très

différente de l'époque couverte par la Librairie de PALM. En outre, il s'agit de textes littéraires très différents de nos textes politiques.

Pour le latin, la recherche est plus avancée. Il existe des dictionnaires électroniques pour le latin classique, ainsi qu'un DELAF créé à partir du lexique Du Cange par une équipe de l'École des Chartes. Il existe également des taggeurs qui ont été entrainés sur les sources antiques, développé à l'université de Liège par le LASLA. D'autres taggeurs ont été développés ou sont en voie de développement à l'École des Chartes (Omnia) et l'université de Francfort (Historical Semantics Corpus Management).

Pour le moyen anglais, nous avons pu établir une collaboration avec le *Linguistic Atlas of Early Middle English* (LAEME) basé à l'université d'Édinbourg. Cette équipe travaille depuis la fin des années 1980 sur l'annotation de textes en moyen anglais, et ils ont créé, entre autres, un système d'annotation pour leurs propres besoins, notamment pour pouvoir analyser la variation orthographique entre les dialectes. Nous travaillons actuellement sur la conversion de leurs formes annotées en dictionnaire électronique pour l'utiliser dans la lemmatisation.

La prochaine étape : L'annotation morpho-syntactique

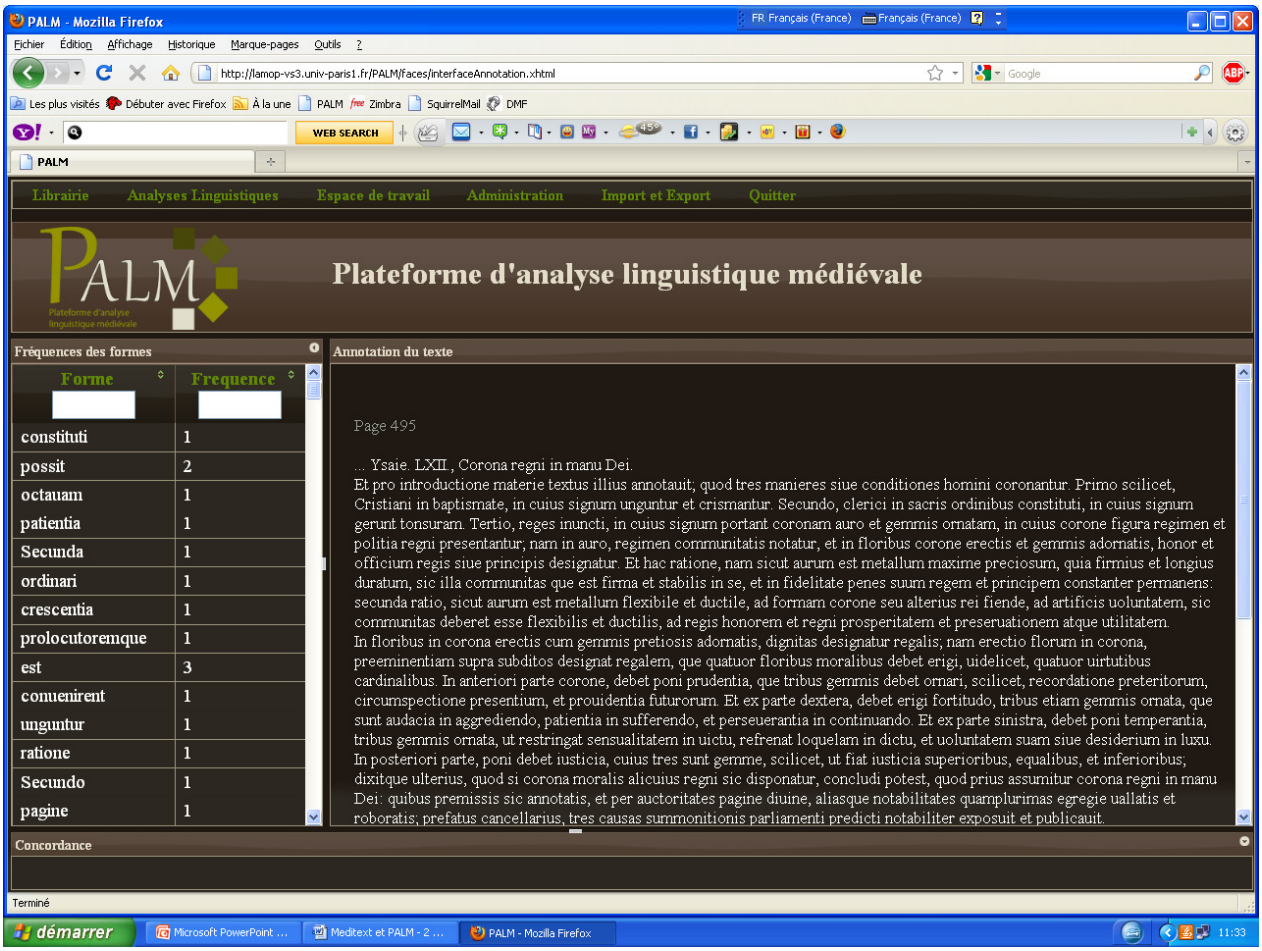

Fig 8: L'annotation morpho-syntactique : Un sermon de John Stafford, 1437

Pour le moment, après avoir développé un système de gestion de corpus, nous travaillons surtout sur l'adaptation et la création des outils informatisés de base pour faciliter la lemmatisation semi-automatique. Ainsi avançons-nous vers la prochaine étape : l'annotation de nos textes et l'entraînement de taggeurs.

Nous avons déjà mis en place l'architecture dans laquelle l'annotation prendra place. Par exemple, si l'utilisateur prend un texte de son corpus, un sermon d'ouverture du Parlement en 1436 pendant le siège de Calais, prononcé par le chancelier John Stafford, il peut cliquer sur « Analyse morpho-syntaxique » pour commencer l'annotation.

La plateforme présente d'abord le texte en deux fenêtres : le texte et une liste de toutes les formes du texte et leurs fréquences [fig. 8]. Si l'utilisateur clique sur une forme dans le texte – comme ici « regni » [fig. 9] – une fenêtre apparaît permettant de la lemmatiser. On peut également créer une concordance à partir de cette forme. Il est possible d'annoter à partir de la liste des formes, une fois que la concordance a convaincu l'utilisateur qu'il n'y a qu'un seul lemme pour cette forme. Sinon, on peut procéder une forme après l'autre, à partir du texte.

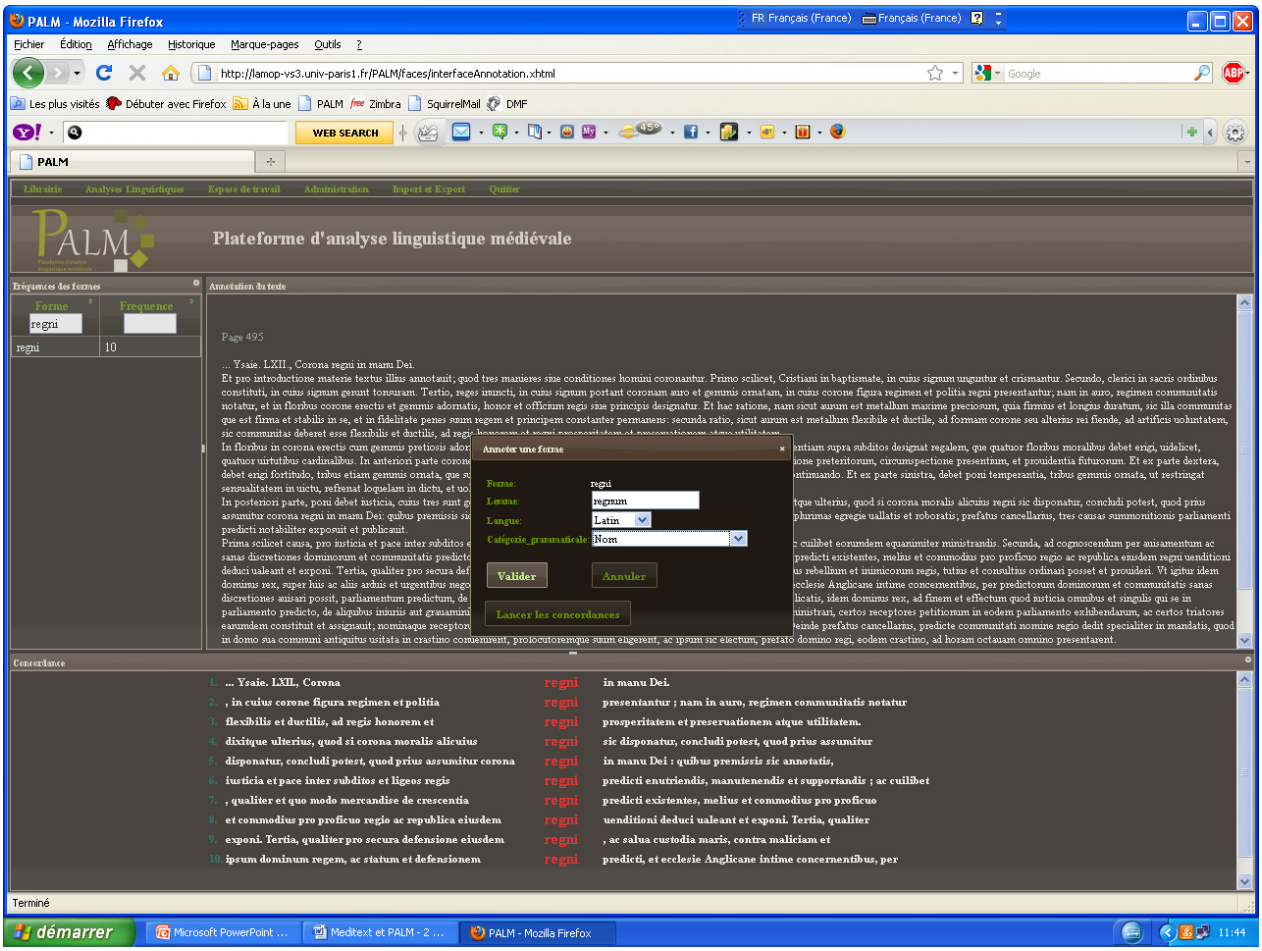

Fig 9: Création d'une concordance ; l'annotation d'une forme.

Il est donc déjà possible de lemmatiser tout un texte à partir de cette plateforme. Toutefois, bien que cela soit plus aisé que la lemmatisation « à la main », cela reste long. Nous l'avons vu, nous sommes actuellement en train de passer à la prochaine étape : l'annotation des textes de notre corpus dans le but d'entraîner des « taggeurs ».

Dans le passé, la lemmatisation d'un texte ne servait qu'à l'étude en cours. Les données étaient perdues à la fin. Mais avec PALM, nous allons annoter les textes de notre corpus et conserver les annotations dans notre librairie. Ces textes annotés seront ensuite utilisés pour entraîner des « taggeurs ». À l'avenir, lorsqu'un nouveau texte sera implémenté dans PALM, ces taggeurs lui seront appliqués.

Ils feront au début beaucoup d'erreurs. Mais en corrigeant ces annotations et en utilisant la plateforme d'annotation au sein de PALM, de nouveaux textes annotés seront créés, qui seront à leur tour utilisés pour entraîner les taggeurs. Ainsi arrivera-t-on à une lemmatisation de plus en plus précise.

État des lieux

Pour l'heure, PALM offre un système de gestion de corpus qui facilite la composition de ces derniers, soit à partir des textes de la librairie de textes politiques intégrée à PALM – par exemple, des traités en latin, des sermons, des discours et des proclamation royales, ou également des lettres ouvertes, des requêtes et de la poésie politique ou morale – soit à partir de textes fournis par l'utilisateur lui-même, soit sur la base d'un mélange des deux. Le prototype de PALM fournit également une interface d'annotation, que nous utilisons actuellement pour créer des nouvelles ressources pour la lemmatisation semi-automatique du moyen anglais, du moyen français et du latin médiéval. Nous avons commencé le travail de création de dictionnaires électroniques à partir des outils dont nous avons pu bénéficier grâce à des collaborations externes et nous élaborons en même temps nos propres ressources. La prochaine étape est l'entraînement des taggeurs sur nos textes annotés. Nous avons déjà créé un prototype pour l'interface d'exportation de la plateforme, qui permet l'exportation de textes bruts et nous sommes actuellement en train de développer des formats pour Lexico 3 et Hyperbase ainsi que pour TXM, un nouveau logiciel en développement sous la direction de S. Heiden à l'ENS Lyon.