

PALM-Méditext

A web-based platform to enable the linguistic analysis of medieval texts (PALM)
A corpus of political texts from late medieval England and France (Méditext)

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

PALM (Plateforme d'Analyse Linguistique Médiévale) is a software utility, operated over the internet, which makes possible the computer-aided analysis of medieval texts. This initial version is focused on political texts of French and English origin, in Medieval Latin, Middle French and Middle English, dating from the twelfth to the early sixteenth century, but with the possibility of expansion to consider other themes, languages and regions. Before PALM, it was impossible, or at least very labour intensive, to apply such an analysis to a large body of texts from this period, as a result of the absent of standard spelling in these languages at this time and (especially in the case of Middle English) the still unformed nature of their grammar and syntax.

PALM is just one part of a much larger programme directed by J.-P. Genet, under the title *Signs and States*. The general aim of this ERC project is to examine the multiple ways in which the development and application in the later middle ages of various forms of rule, of state power, or government, in the later middle ages changed societies not only socially or economically, but also culturally. Thus the main focus of the program is on symbolic power.

Historians tend to underline the transformative role played by two central factors in the formation of the late medieval 'state': the development of princely justice and of public taxation. Across Europe, the authority characteristic of kings was increasingly monopolised by single authorities with precise territorial claims. This authority was increasingly promoted at the expense of customary arrangements and local rights, through the use, notably, of arguments drawn from Roman law, to expand princely jurisdiction and justify innovative legislation. For example, arguments of evident necessity, and the elaboration of concepts of representation, permitted kings and princes were able to raise new taxes and to fund armies of unprecedented size. However, this was only possible as a result of the evolution of a cultural context characterised by a shared system of communication which permitted what we can call a 'political society' to develop. The new values linked with the emergence of new state apparatus co-existed with older systems of values sometimes counteracting them, sometimes creating something else entirely.

These are themes which have been considered by historians for many decades now. Most recently, historians have focused on popular political ideas; in 'political culture'; in the concept of the 'public sphere'; in politics conceived not only as high politics but as an interaction at many social levels; and in the nature of the late medieval 'state'. The later theme, in particular, have been pursued in the 1980s and 1990s in a series of conferences and their associated publications funded by the European Science Foundation, and directed by J.-P. Genet. These focused largely, although not exclusively, on the development of taxation, justice and on 'political society' – the constantly changing group of those who played an active role in politics. The present European Research Council project has made possible a further important step. The aim of *Signs and States* is to consider less the development of state phenomena in themselves, and more their impact on

mentalities, and their interaction with existing culture. The aim is to consider the commonplace ideas and elaborated values of the populations who were touched by, and who took part in, the development of these state mechanisms, at the same time as ‘top-down’ innovations.

One approach to these problems is the study of terminology. This study has greatly expanded since the late 1960s in the work of what is sometimes referred to as the ‘Cambridge School’ of the history of political thought, associated notably with P. Laslett, Q. Skinner and J.G.A. Pocock. One should also consider the development over the same period of the *Begriffsgeschichte* associated with R. Koselleck and H. Gumbrecht, which has the particular merit of underlining the political, social and economic contexts in which terminology develops. Nonetheless, until very recently, this work necessarily had to focus on either a single author or group of authors, or a single term or group of terms. In this present program, although the texts of the authors who have been traditionally the basis of these studies are still investigated, the main object of research is more ambitious: language itself as a system of signs which is one of the main vectors through which (in the terms of M. Godelier) the *idéel* of a given society is conveyed. The statistical methods which enable the analysis of the whole vocabulary of a large corpus could only be applied with difficulty without the powerful computing resources only recently available. In the case of the middle ages, the nature of texts without standard spelling has made even straight-forward semantic analysis, the pursuit of a particular term, say, an enormously time-consuming task especially for a single researcher.

This is the purpose of PALM: to enable the computer-aided exploration and especially the semantic and statistical analysis of late medieval political texts.

The following documents present the work of the team dedicated to this aspect of the ERC project *Signs and States: Les vecteurs de l'idéal*.

1. An introduction to PALM
2. A list of the texts in the Library of PALM (Méditext)
3. A technical presentation of PALM
4. International contacts of the team
5. Bibliography of the team
6. Presentation and minutes of the Journée d'Études ‘Meditext’ held on 18-19 May 2011

1. An Introduction to PALM

What is PALM?

It is important to stress that PALM is above all a ‘platform’ : that is to say a utility which brings together existing computer resources for the treatment of medieval texts, whilst at the same time developing new ones. In particular, it is important to underline that PALM does not seek to provide advanced statistical analysis itself. Its function is to convert texts into a form which can be treated by existing software for the computer analysis of texts. We do not intend to ‘reinvent the wheel’. For the moment, PALM is intended to export to applications developed in France, such as Lexico 3, Hyperbase and TXM, but its structure is conceived to be entirely expandable for export towards existing and future utilities.

Before using the most generally available computer applications for textual analysis it is necessary to pass through three steps:

1. the composition of the corpus.
2. ‘lemmatisation’ – the most difficult and labour-intensive stage
3. the preparation of the corpus for treatment by a particular application

PALM is intended to make these three tasks easier, and also, as far as possible, to make them automatic. PALM is web-based, but access to it is secured by a login and password, obtained by application to us. The web address is <http://lamop-vs3.univ-paris1.fr/PALM>

Using PALM

1. Composition of a corpus

First, then, PALM provides an environment in which a corpus can be constituted either from texts which the user brings to the platform, or from PALM’s own ‘Library’ [see fig. 1].

It has been the putting together of the ‘Library’ within PALM which dominated the first year of our work, roughly from October 2010 to October 2011 (after an initial phase spent in recruitment) and the exploration both of possible future texts to include and of existing resources at Paris I.

The basis of PALM’s Library is ‘Meditext’, a corpus of texts of English and French origin which had been assembled piecemeal since the early 80s, either in person by J.-P. Genet and Cl. Gauvard or under their direction.

This corpus is made up of ‘political’ texts, by which we mean:

- either texts related to specific political events
- or texts which consider good or bad rule in a general manner

For the moment, PALM’s library is made up of texts of English origin, in French, Latin or Middle English, or of French origin, in French or Latin. In the future, however, we hope to include texts from different European countries, and the platform has been put together in such a way that this will be possible.

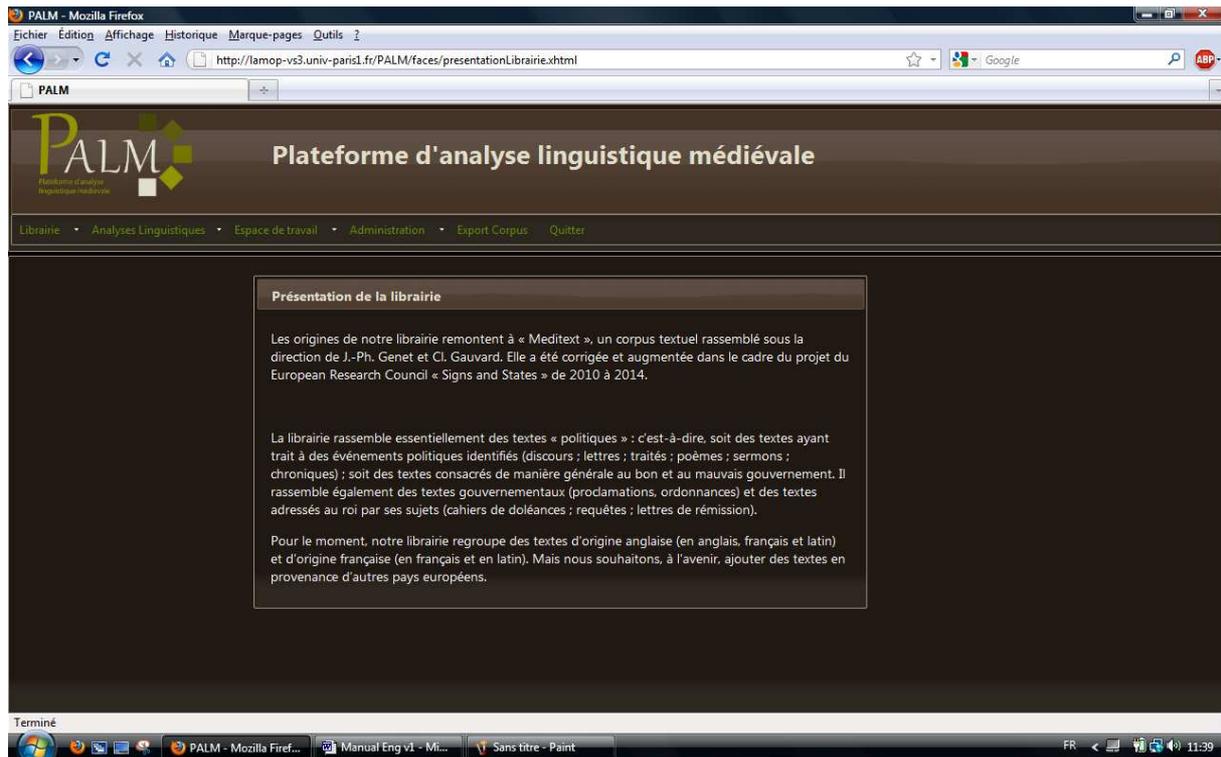


Fig 1: Welcome page after login to PALM: Description of the 'library' within PALM

Méditext and PALM's integrated Library

For a detailed presentation of the texts within PALM, please see the attached annex. Some remarks, however, should help give an impression of the kind of texts with which we are dealing.

Texts of French origin

The texts of French origin are mostly in French, although some of them are in Latin. First of all, we have a number of political treatises, digitised under the supervision of Claude Gauvard, such as the *Quadrilogue Invecitif* by Alain Chartier or the *Songe du vieil pèlerin* by Philippe de Mézières. In addition we also have two corpora of a rather different kind, also digitised under the supervision of Cl. Gauvard. The first is made up of a selection of *ordonnances* issued by Charles VI. The second is a corpus of letters of remission – documents granting a royal pardon which reproduce the petitions in reply to which these were given, which tell the stories of the petitioners in often considerable detail. These, again are mostly from the late fourteenth or early fifteenth century.

Initially, these texts were in a variety of different formats characteristic of the period when they were digitised. The earliest had even been inputted line by line on punch cards, and then scanned to preserve them. This might at first seem a waste of effort, until it is realised that modern text-recognition software (OCR) works very badly for texts in medieval languages (they do not possess standard dictionaries by which they can correct them) and in older editions (they have difficulty recognising nineteenth century fonts and formatting). The work of converting these files into a form which could be used took up most of the time of our Middle French linguist, N. Kanaoka, from her recruitment in October 2010, up until the end of 2011.

Recently, however, she has been able to expand this corpus, starting work on a corpus of political treatises and political poems in French. She has notably expanded our corpus by correcting externally commissioned transcriptions of the French version of the *De Regimine Principum* of Giles of Rome, a political treatise which was extremely widely circulated in the later middle ages. We have also received a large corpus of the sermons of Jean Gerson. It is our intention to add a further corpus of historical texts, notably the *Grandes Chroniques de France*, and selections from Froissart's *Chroniques*.

Texts of English Origin

It should be stressed that the texts of English origin are in three different languages: French, Middle English and Latin. From the Norman Conquest until the end of the twelfth century, written practice in England took place overwhelmingly in Latin. From the mid-twelfth century, French began to emerge as a literary language and, by the end of the thirteenth century, as the language in which it seemed appropriate to represent legal proceedings, letters, public proclamations and petitions. Ironically, this shift to French occurred during precisely the period that this language was waning as a widely spoken second language amongst the English nobility. From the end of the thirteenth century, English begins to emerge, first as an appropriate medium for religious instruction or popular literary entertainment, but increasingly, from the end of the fourteenth century, as a medium for record keeping on a local level, more elaborate literary works and even letters, petitions and political treatises. Still, even at the end of the middle ages, French, English and Latin, continued to be used, sometimes in the same document, by literate readers and writers in England.

The texts in our corpus of English origin fall mainly into three groups:

- (1) On the one hand, there is a corpus of reports of speeches, short treatises, royal proclamations and letters associated with specific political events from the thirteenth to the sixteenth century, found in royal and private archives, in all three languages, notably texts concerning Parliament assembled by Jean-Philippe Genet.
- (2) On the other hand, there is a large corpus of political poetry, once again linked to specific political moments, largely in English, collected by J.-P. Genet and by Aude Mairey.
- (3) This corpus is rounded off by a number of sermons and political treatises in English, from the fourteenth to the sixteenth century.

This corpus has been expanded by our Middle English linguists, for the first year Rachel Moss, and as of November 2011, by Chloe Morgan. We have added texts in English, notably the English translation of Giles of Rome by John Trevisa, a number of fifteenth century political texts taken from the commonplace book of John Vale, and now a number of historical texts, such as the English version of the *Brut* chronicle, and the translation, again by John Trevisa, of the widely circulated *Polychronicon* of Ranulph Higden. At the same time, we have added political poetry of English origin in Latin and French from the thirteenth and fourteenth centuries, and a number of texts of a political kind, issued by the royal government or addressed to it: royal proclamations, open letters, and again ordinances and political treatises.

Some remarks on Latin texts

At the start of this project, we had considerably fewer texts in Latin than in French or English, and the texts we did have were above primarily either English royal proclamations, letters and speeches, or sermons to be found in the records of the English Parliament. Aside from these, we also had the chronicle of the monk of Saint Denis. Moreover, work on Latin texts was initially held up by some recruitment problems. Our first Latinist, a thesis student, thus started only in January 2011 and stayed for one year.

During this period, though, besides checking the material in Méditext and making it ready for PALM, she also added an important political treatise, the *Policraticus* of John of Salisbury; a number of political poems in Latin; and several political texts linked, for example, to the crises of the first half of the fourteenth century in England. The digitisation of the Latin *De Regimine Principum* of Giles of Rome was undertaken externally thanks to the SAS project which financed it. Work has begun on correcting it, alongside a digitisation of the *De Regno* of Thomas d'Aquin undertaken at the university of Nancy.

From January 2012, we have been joined by Laura Albiero. Besides assisting with the creation and adaptation of linguistic resources (on which more in a moment) she has begun to add historical texts, notably the Latin version of the *Polychronicon*, and a number of political treatises of European import, notably these linked to the controversy between Philip IV of France and Pope Boniface VIII.

How do you create a corpus using PALM?

The first task of our software engineer, Mourad Aouini, was to construct an environment for the consultation and manipulation of this 'Library' in Middle English, Middle French and Latin. The technical aspects of this task are described in the attached annex.

Once the user has logged into PALM, she can then browse through the texts present in the 'Library' [see fig. 2]. She can search the texts by keyword in each of the field, or by alphabetical order: say by language, or by country of origin.

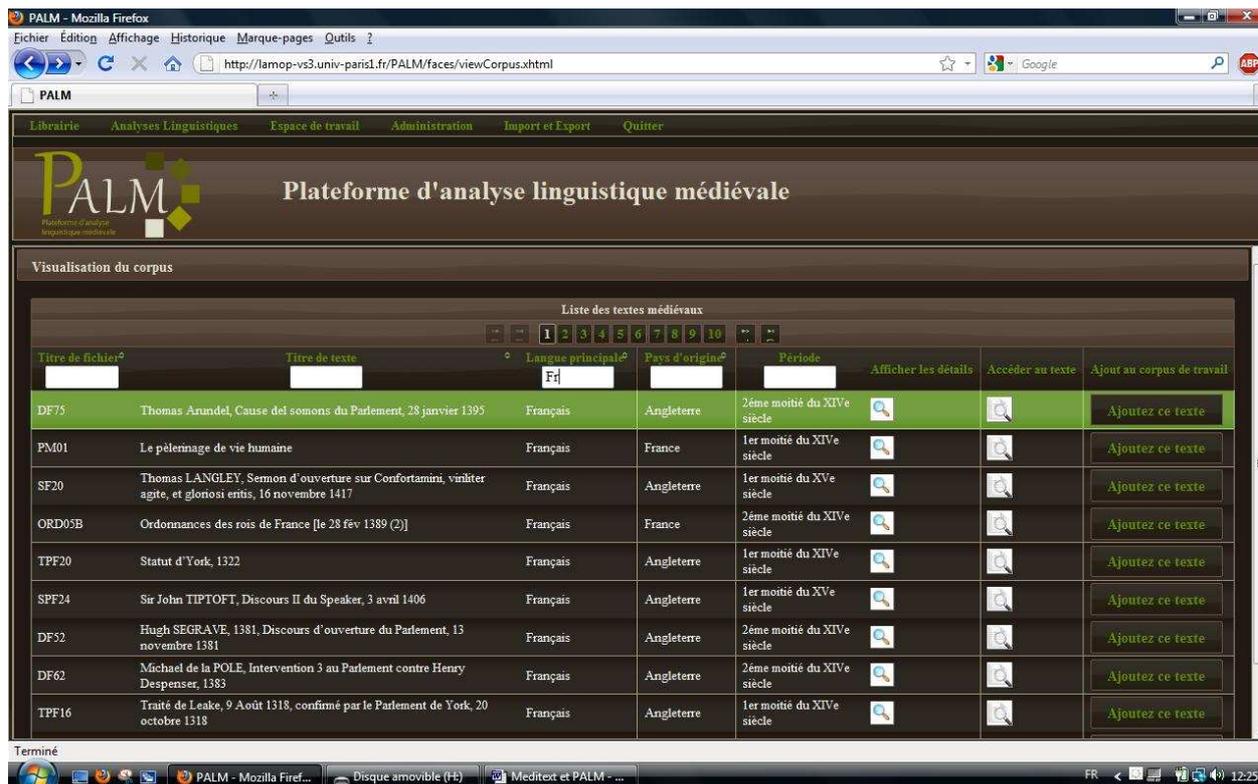


Fig. 2: The 'Library' organised by main language: French.

You can also consult a brief description of each text – so here, for example, we have three political ballads by the late fourteenth and early fifteenth century writer, Jean Creton.

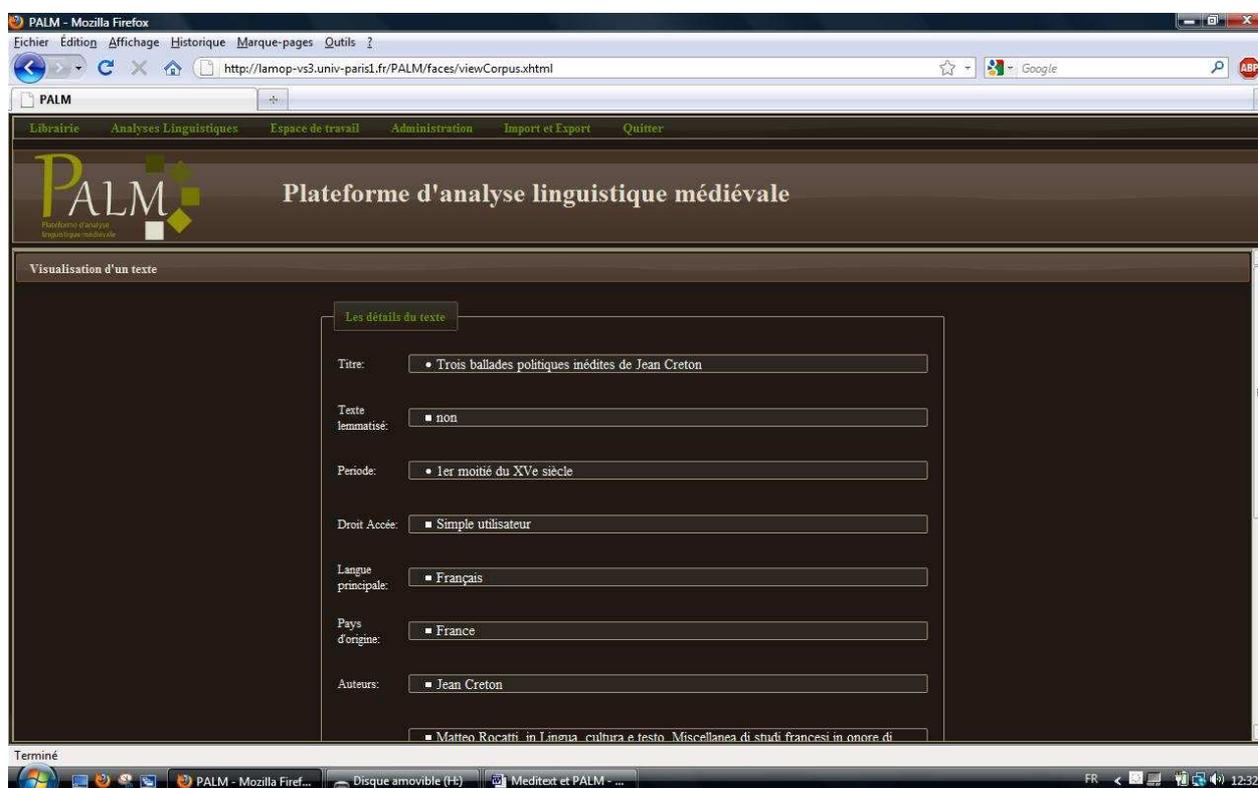


Fig. 3: Details: Three poems by Jean Creton

It is also possible to consult the text [fig. 4].

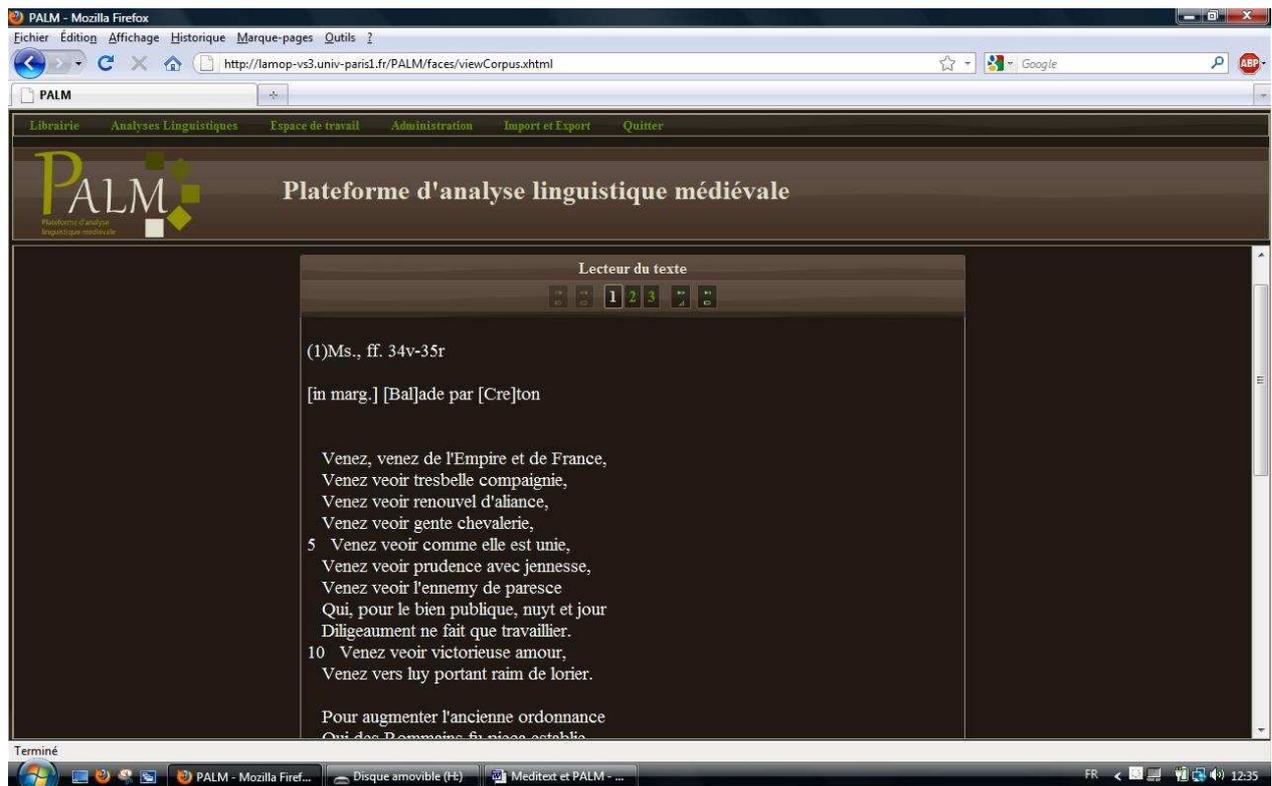


Fig 4: *The Library: Three ballads by Jean Creton: Consultation of the text*

So, here we are consulting texts which are in the central Library of PALM. But, at each step, the user can decide to select a text and place it in her own personal corpus – what we call her ‘Corpus de travail’ or ‘Workspace’. If she wanted to, the user could construct your corpus just on the basis of the texts in the Library. But it is also possible to import external texts to her personal ‘Corpus de travail’. For the moment, this is done by using a form, found in the menu for the library ‘Ajouter un texte’ (‘Add a text’).

So, for example, imagine that the user wants to analyse a sample of petitions submitted to the king by the Commons in the English parliament of January 1437 [fig. 5]. First of all, she enters the details concerning her text’s origin – when and where it was composed, and which edition the text comes from, a long and a short title, and so forth.

She can also set its level of access: whether it will be available to any user, whether it will be on restricted access, or whether it will only be available to the administrator – for example in the case of editions currently in progress.

In the version from which this slide is taken, the text is added by cutting and pasting it into the window “page” – but a feature has now been added by means of which you can download from a text file or word.

It is also possible to add pagination with flags in the style <p=1>.

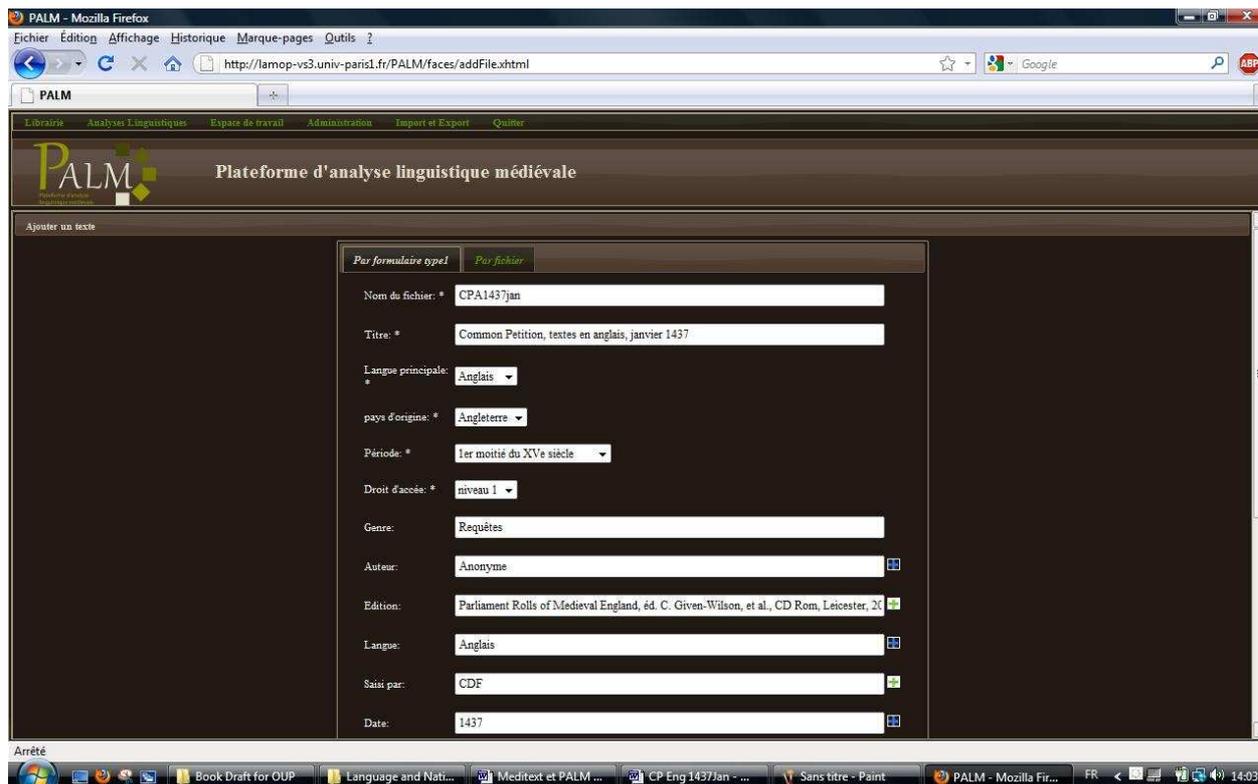


Fig. 5: Adding a new text to the user's personal Corpus de travail

If the user clicks on 'Envoyer le texte', her text is downloaded into her workspace. She can then consult it in the list of texts by clicking on 'Corpus de travail'. She can look at its details or the text itself.

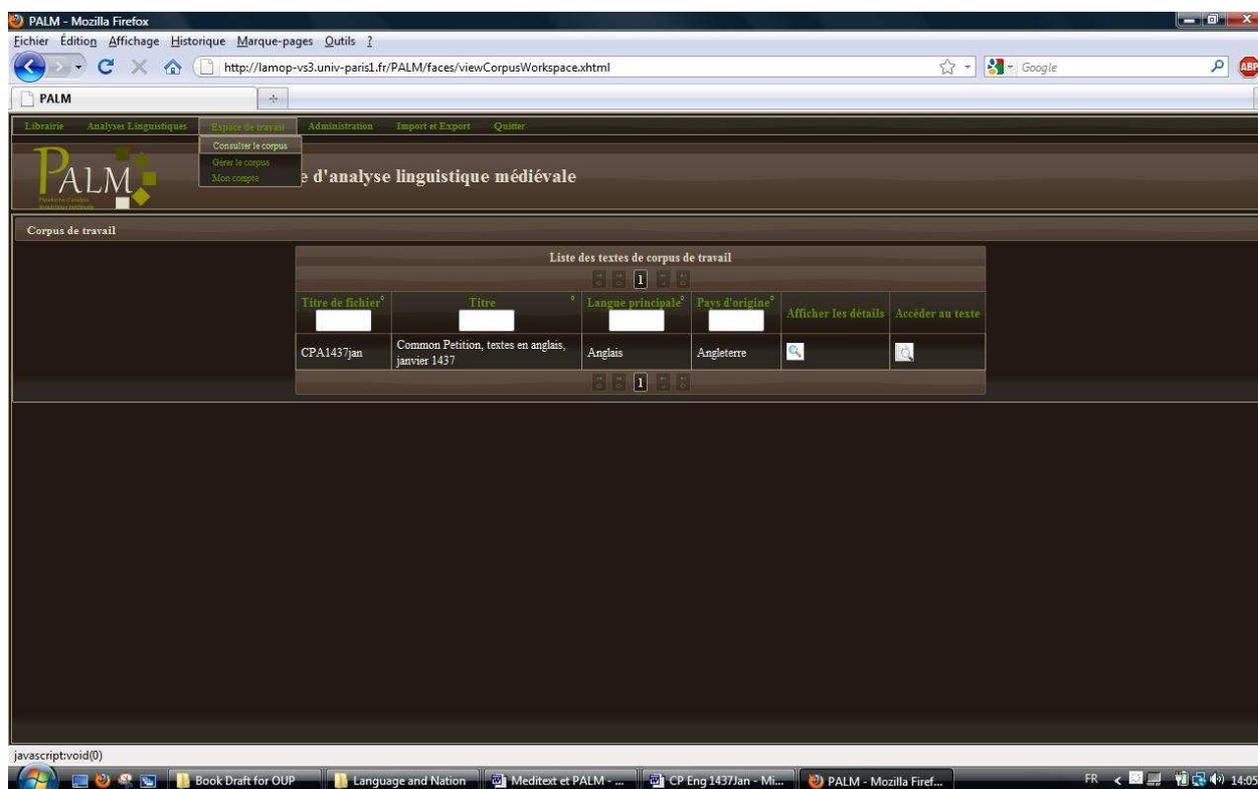


Fig 6: Corpus de travail with a single text added.

Now, suppose the user wants to compare the language of this text with some others from the same period, just after the treaty between the king of France and the duke of Burgundy in 1435, and the siege of English-occupied Calais by Philip the Good, duke of Burgundy, in 1436. She can go back to the central library, select the relevant texts and click on ‘Ajouter ce texte’ in each case. They appear in her workspace, ready to pass to the next stage.

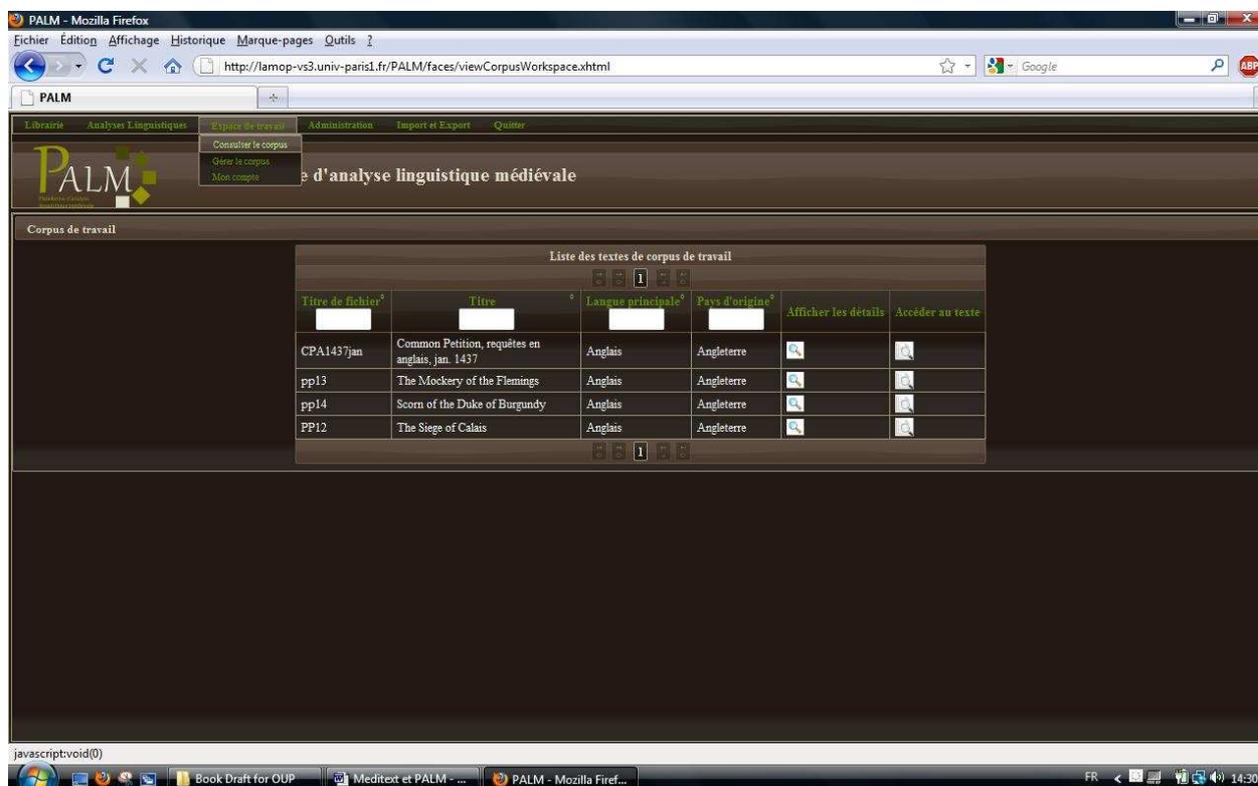


Fig. 7: A small corpus ready for lemmatisation

Through the menu ‘Espace de travail’, the user can modify the details of her texts, or remove them from her corpus. Through the menu ‘Administration’, she can change her password and personal details.

Already, then, the platform makes the composition of an experimental corpus easier, and provides access to the texts in the library. That, then, is step one. This is, however, by no means the most difficult step. That comes next: the lemmatisation stage.

2. Lemmatisation

In order to use most computer applications which provide linguistic analysis, it is essential, either to get good statistical results, or even to perform efficient searches, to lemmatise the texts to be analysed. In this case, lemmatisation means annotating each word in our texts with its ‘lemme’, which is to say the canonical form of the word which would be found in a dictionary of that language.

The benefits of lemmatisation vary according to the language under consideration, and the historical period from which it is drawn. In the case of modern French, for example, the primary advantage is to group together different inflected forms of nouns, verbs and adjectives. So, for example, *vous aimez*, *nous aimons* and *tu as aimé* are all annotated with the infinitive ‘aimer’.

The noun in *deux chevaux* or *un cheval* are both linked to the dictionary form *cheval*. Lemmatisation in any inflected language thus makes it possible to consider at the same time all the forms of a verb or a noun, for example.

In medieval languages, though, especially medieval vernacular languages, lemmatisation is even more important, but also even more difficult, because of the absence of standard spelling. So, to stay with French for the moment, if one consults the online *Dictionnaire du Moyen Français*, which deals with French between the thirteenth and the fifteenth centuries, you will find many different spellings for even the simplest words. So, in the relatively simple case of ‘aimer’, for example, the DMF, gives three variants: ‘amer’, ‘aimer’ or ‘aymer’. This though is just for the infinitive, and the variants are increased by the inflections which can be derived from it.

The longer the word, the greater the possible variation. So, for ‘léopard’ – leopard – the DMF gives ‘liepart’, ‘lieppart’, ‘leoppart’, ‘lupar’ and ‘lupar’ in the singular, and ‘liepars’, ‘lieppars’, ‘limpars’ and ‘luppars’ in the plural. Even for the ‘seigneur’, lord, which is very frequent in our texts, the DMF gives nine possible spellings in the singular, five in the plural: *sainieur, segneur, seigneur, seignour, seigny, seingneur, siegneur, signeur* and *singneur*; *saingnours, scenours, segneurs, seigneurs* and *seigners*.

The situation in Middle English is, in fact, even more difficult. So, if we take just one text, the *Ayenbite of Inmyt*, a religious text from the second quarter of the C14, and extract all the forms noted in the annotated texts provided by the *Linguistic Atlas of Early Middle English* (LAEME) we get, just for the word for ‘kued’ a form which in modern English gives ‘wicked’, 10 forms: *kuead, kueade, keead, quead, kued, queade, kuade, kuede, kueades, kuedes*. Or, in the same text, ‘saint’, 16 forms: *saynt, zaynte, saint, sayn, zayte, seynt, zainte, sanyn, sanyt, saynte, seint, zainte, zaynte, zayn, sayn* and *saunynt*.

Moreover, if we take the example of the reflexive pronoun, ‘himself’, which could also simply take the form ‘him’, and consult all the LAEME annotated texts provide around 90 forms for just one lemme: *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, be-seolf, him ysself, him-selfen, im-self, him-seoluen, hine-selue, him-selfi, him-sulfne, him-suelf, him-sulne, hine-seolfne, hine-seolfne, hine-seolf, him-solf, hm-solf, him-seelf, hine-sulfne, hine-seulfne, hine-sulue, him-suluen, himseoluen, him-seluein, hym-selue, him-seluen, him-sulfen, him-selge, 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-seolfne, himsuluen, him-seolue, hire-seolue, him-sulf, him-suluen, himm-sellfenn, himseluen, hym-sylfe, hine-silfne, him, himm, hym, im, hem, hine, hym, hyne, hine*.

As things currently stand, in order to analyse texts in Middle English or Middle French statistically it is necessary to lemmatise them by hand. Even in order to search them, it is necessary to pick up the particular forms of a given lemme and search for them one by one.

As can be imagined, replacing all the forms in a text with their respective lemma, or even with a standardised spelling, is always an enormous task. Normally, the historian has no choice but to restrict herself to the particular words which interest her, which makes the results of the final analysis less significant.

The main purpose of PALM is thus: first, to bring together in a single utility all the available resources which facilitate the lemmatisation of a particular text; and second, in the process, to create new linguistic resources.

Linguistic Resources

At the current stage in the project, the lemmatisation stage is the primary focus of our work. The main problem is that there are simply not the same linguistic resources available for the computer analysis and lemmatisation of late medieval English, French and Latin as there are for modern languages. For the moment, we have focused on two types of linguistic resources which are useful in the process of lemmatisation. It is helpful to introduce them briefly here, although they are considered in more detail in the technical annex.

a) Electronic Dictionaries

For modern languages, there exist readily available ‘electronic dictionaries’, in the sense of lists of forms, annotated, for example by lemma and grammatical category. In French, it is common to talk of DELAFs or Dictionnaire ELectronique de Formes fléchies – Electronic dictionaries of inflected forms. By applying a dictionary like this to a modern language, a computer can suggest different possible annotations for a given example.

Nonetheless, electronic dictionaries are just one resource, with their own limits. If one applies an electronic dictionary to a text one word after another, without taking into account the grammatical context, it cannot resolve ambiguities which pose no problems for a human reader. So, for example, many forms could refer to nouns or verbs, but this normally presents no problem. For example, in modern French, one might take ‘Il a marché sur la lune’ or ‘Il est allé au marché’; or in modern English: ‘Don’t step on my toe!’ or ‘He swept the front step’. A human reader has no difficulty in distinguishing between the verb and the noun with the same form, but a form dictionary, which treats a sentence word by word and not sentence by sentence, cannot do so.

So, in this case, another approach is necessary. One method is to make use of a ‘tagger’.

b) Taggers

Taggers are applications based on artificial intelligence. They can be used to suggest solutions to ambiguous cases thanks to a preliminary ‘training’ on texts which have been annotated by human users. They can then suggest the most probable answer in cases where a dictionary alone produces ambiguity.

But at this point we reach the bad news. There exists no ‘electronic dictionary’ for Middle English and, as far as we have been able to discover, nobody has ever trained a tagger on this language. This situation is all the more serious in that, as we have seen, variation in spelling in Middle English is very marked. Moreover, the grammar of Middle English is still a long way from the standardised grammar even of, for example, seventeenth or eighteenth century English.

The situation is a little less serious in the case of Middle French. The team at the *Dictionnaire du Moyen Français* at Nancy have created a utility (LGerM) which can suggest a number of solutions to a user with an unknown form. Nonetheless, this tool does not lemmatise automatically. The user still has to choose amongst numerous suggestions. But it is possible to apply it to certain texts through it, and then sift the list of suggestions to create one’s own DELAF. Other teams have trained taggers on Old French, but unfortunately this is very different from Middle French, and the texts in question are of a very different kind from our political texts.

For the Latin language research is slightly more advanced. There are electronic dictionaries available for classical Latin, and we have also been able to use the dictionary created on the basis of the Du Cange lexicon by the team at the Ecole des Chartes. The Historical Semantics Corpus Management system under development at the university of Frankfurt also permits the lemmatisation of Latin texts through the application of a tagger. Taggers have been trained on classical sources by at the university of Liège as part of the LASLA research project, and a further tagger has been trained on earlier medieval sources at the École des Chartes (Omnia).

Creating an Electronic Form Dictionary can only be the fruit of many years of work. For Middle English, we have been able to establish a collaboration with the Linguistic Atlas of Middle English at the university of Edinburgh. This team has been working on the annotation of English texts from the thirteenth and fourteenth centuries since the late 1980s. For their own reasons, they prepared a system of annotation, aimed first of all at tracing dialect variations, but which in the end provides a very useful resource of annotated texts. We are currently working on the conversion of their annotated forms into an Electronic Form Dictionary which can be used in the process of lemmatisation.

The Next Step: Morpho-syntactic Annotation

The screenshot shows the PALM web interface in Mozilla Firefox. The browser address bar shows the URL: <http://lanop-vs3.univ-paris1.fr/PALM/faces/interfaceAnnotation.xhtml>. The interface has a navigation menu with items: Librairie, Analyses Linguistiques, Espace de travail, Administration, Import et Export, and Quitter. Below the menu is the PALM logo and the title "Plateforme d'analyse linguistique médiévale".

The main content area is divided into two panels. The left panel, titled "Fréquences des formes", contains a table with two columns: "Forme" and "Frequence".

Forme	Frequence
constituti	1
possit	2
octauam	1
patientia	1
Secunda	1
ordinari	1
crescentia	1
prolocutoreque	1
est	3
conuenirent	1
unguntur	1
ratione	1
Secundo	1
pagine	1

The right panel, titled "Annotation du texte", shows a text editor with the following content:

Page 495
 ... Ysaie LXII, Corona regni in manu Dei.
 Et pro introductione materie textus illius annotauit, quod tres manieres siue conditiones homini coronantur. Primo scilicet, Christiani in baptisate, in cuius signum unguuntur et crismantur. Secundo, clerici in sacris ordinibus constituti, in cuius signum gerunt tonsuram. Tertio, reges inuinci, in cuius signum portant coronam auro et gemmis ornatam, in cuius corone figura regimem et politia regni presentantur, nam in auro, regimen communitatis notatur, et in floribus corone erectis et gemmis adornatis, honor et officium regis siue principis designatur. Et hac ratione, nam sicut aurum est metallum maxime preciosum, quia firmitus et longius duratum, sic illa communitas que est firma et stabilis in se, et in fidelitate penes suum regem et principem constanter permanens: secunda ratio, sicut aurum est metallum flexibile et ductile, ad formam corone seu alterius rei fiende, ad artificis uoluntatem, sic communitas deberet esse flexibilis et ductilis, ad regis honorem et regni prosperitatem et preseruacionem atque utilitatem. In floribus in corona erectis cum gemmis pretiosis adornatis, dignitas designatur regalis; nam erectio florum in corona, preminentiam supra subditos designat regalem, que quatuor floribus moralibus debet erigi, uidelicet, quatuor uirtutibus cardinalibus. In anteriori parte corone, debet poni prudentia, que tribus gemmis debet ornari, scilicet, recordatione preteritorum, circumspectione presentium, et prouidentia futurorum. Et ex parte dextera, debet erigi fortitudo, tribus etiam gemmis ornata, que sunt audacia in aggreddiendo, patientia in sufferendo, et perseuerantia in continuando. Et ex parte sinistra, debet poni temperantia, tribus gemmis ornata, ut restringat sensualitatem in uictu, refrenat loquelam in dictu, et uoluntatem suam siue desiderium in luxu. In posteriori parte, poni debet iusticia, cuius tres sunt gemme, scilicet, ut fiat iusticia superioribus, equalibus, et inferioribus; dixitque ulterius, quod si corona moralis alicuius regni sic disponatur, concludi potest, quod prius assumitur corona regni in manu Dei: quibus premissis sic annotatis, et per auctoritates pagine diuine, aliasque notabilitates quamplurimas egregie uallatis et roboratis, prefatus cancellarius, tres causas surmonitionis parlamenti predicti notabiliter exposuit et publicauit.

At the bottom of the interface, there is a "Concordance" search bar and a "Terminé" status indicator. The Windows taskbar at the bottom shows the system tray with the time 11:33 and several open applications including Microsoft PowerPoint and Meditext.

Fig 8: Morpho-syntactic annotation: A sermon of 1437 by John Stafford

For the moment, having developed a corpus management system, we are working above all setting up some basic linguistic resources necessary for semi-automatic lemmatisation within PALM. That will allow us to move forward on the third step: the annotation of our corpus, and then the training of taggers on them.

We have already set up the architecture of within which this process will take place. So, if the user takes a text from her corpus – a sermon for the opening of Parliament in 1436, during the siege of Calais, delivered by the chancellor, John Stafford. We can then click on ‘Analyse morpho-syntaxique’ and begin annotation.

PALM presents this text first of all in two windows: the text; a list of all the forms in the text and their frequencies [fig. 8]. If the user clicks on a word in the text – such as ‘regni’ here [fig. 9] – a window pops up which allows me to lemmatise it. The user can also create a concordance based on that form. If the user wishes to, she can annotate on the basis of the list of forms, once the concordance has convinced me that we are dealing with only a single lemme for this form, or precede form by form.

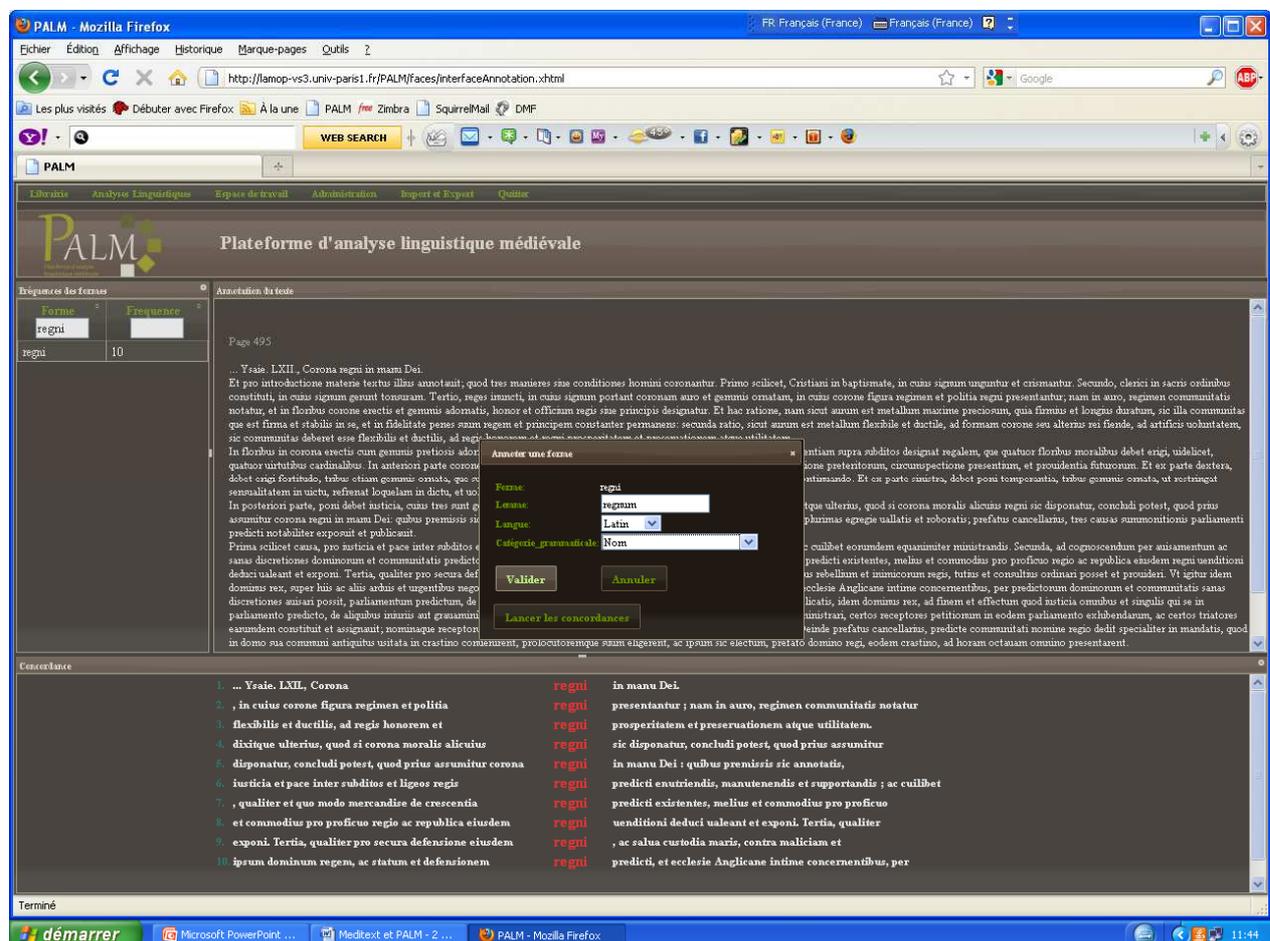


Fig 9: Creating a concordance; annotating a form.

Already, it is possible to lemmatise an entire text using this platform. Nonetheless, although this is rather more practical than manual lemmatisation, it is still a long process. As has already been suggested, we are currently working on the next step: the annotation of texts in our corpus in order to train taggers.

In the past, if a text was lemmatised, the lemmatisation was only useful for the purpose of the project in hand. The data was lost at the end. But within PALM, annotations on the ‘Library’, or new texts which are introduced, will be kept. These annotated texts can then be used to train ‘taggers’. In the future, when a new text is put into PALM, these taggers will be applied.

At first these will produce high levels of error. But by correcting these annotations, using the annotation facilities provided by PALM, we will create new annotated texts, which will be used in turn to train taggers. In this way, the lemmatisation will get more and more precise.

Prospects

At this mid-project stage PALM provides a corpus management system which enables the composition of corpora, either on the basis of its considerable integrated library of political texts – ranging from learned Latin treatises, sermons, speeches and royal proclamations to opposition treatises, petitions and works of political and moral poetry – and/or on the basis of uploaded texts provided by the user. The prototype also provides an interface for annotation, which we are using to build resources for the semi-automatic lemmatisation of Middle English, Middle French and Medieval Latin. In recent months, we have begun the work of compiling electronic dictionaries of these languages on the basis of materials received from external collaborations as well as those developed by ourselves. The next step is to train taggers on our annotated texts. The architecture for export is also present in prototype form, enabling export in simple text format, and we are currently developing export formats for use with Lexico 3 and Hyperbase, alongside collaborations to enable export into TXM, a new utility under development by S. Heiden at ENS Lyon.