VariaLog : How to Locate Words in a French Renaissance Virtual Library

Marie-Hélène Lay

[1] *University of Poitiers. Linguistics Department, France*

Abstract: *The efficiency of search engines is based on the principle that the information sought can be retrieved by "looking for words" conveying the information. This amounts to taking for granted that words are always written in the same way. This view, which is well adapted to texts produced in contemporary periods of language history, is not suited to texts produced during the French Renaissance. The choices made at the Center for Renaissance Studies of the University of Tours, for the Virtual Humanistic Library Project are the subject of the present paper. After due consideration of the strategies based on text annotation, a new tool to extend queries is being put forward to be integrated to an XTF (eXtensible Text Framework) platform. Another purpose of the paper is to show the relevance of resorting to linguistic expertise in order to generate the forms to be sought in texts.*

Keywords: *French Renaissance, Search engine, Query expansion, Virtual library.*

I. SEARCHING WORD FORMS IN A PRE-ORTHOGRAPHIC CONTEXT

The efficiency of search engines is based on the principle that the information sought can be retrieved by "looking for words" conveying the information and that these words can be identified thanks to the string of characters they are comprised of. This view takes for granted that the words are always spelt in the same way and that they comply with orthographic rules.

This view may well be suitable for texts which are produced in contemporary stages of language history, and which correspond to the vast majority of texts available in digitized form. Such is not the situation which prevails for the texts produced during the French Renaissance period. Therefore the availability of older texts for purposes of archiving and disseminating the cultural heritage tradition raises a particular problem.

Texts edited in French before the 18th century are characterized by an irregularity in spelling which raises obstacles in the efficient use of search engines: spellings are not consistent, as proper spelling has not been "invented" yet. One and the same word may therefore be spelt in a variety of forms. This is not only a time-related variation, as would be expected from the evolution of the language between the 15th and the 17th century, but in one and the same book many different spellings may be identified for the one and the same word. For example, one may find the use of either *un* or *ung* in the same text. Similarly for the word *côté*, either *coté, cotté, cote, costé,* or *couste could be used*. The verb *savoir* may be spelt either *scavoir* or *sçavoir*, «*je sais* » may be spelt « *ie sçay* », and its past participle « *su* » may appear as « *sceu* ».

It is therefore necessary to adapt search engines based on word form identification if they are to render the service expected. Several strategies can be envisaged and the purpose of this paper is to focus on those which resort to linguistic expertise. The solutions considered are produced in the particular context of the Virtual Humanistic Library Project and its evolution. The part of the project presented here, VariaLog, is financed by a Google Digital Humanities Research Award.

A. Books and tools

The Bibliothèques Virtuelles Humanistes (BVH), i.e. the Virtual Humanistic Library project (VHL), run by a research unit at the Centre d'Etudes Supérieures de la Renaissance (Center for Renaissance Studies) at the University of Tours, France. This virtual environment aims to disseminate the cultural heritage of the French Renaissance period.

The digitization project, which was started in 2003, (http://www.bvh.univ-tours.fr; Demonet & Lay, 2008) is the continuation of an editorial project (Demonet, 1995), in which the text was already integrated to a lexicometric environment called Hyperbase (Demonet, 1996,), which results in using new tools, available in an electronic environment, to improve the reading and interpretation of literature.

The BVH/VHL project offers two types of digital representations: the image of a copy (its "facsimile"), and its corresponding transcription; 483 books or manuscripts of the Renaissance period (out of a total of 700 digitized works), and 31 transcribed texts are currently online. Several tools are available on the BVH/VHL website, such as the XTF search engine (http://xtf.cdlib.org), using which the reader can access the text as he would in a physical library, by finding the book and then read it. He can also enjoy the possibility of accessing the contents of the book, viz. linguistic information (www.artamene.org/philologic.php), or graphic information (http://www.iconclass.nl/). In order to achieve its goal, the BVH/VHL project also develops its own tools: (1) AGORA (a graphic analyzer for rare books: Ramel, 2006); (2) RETRO project, (dealing with Optical Character Recognition for old printed texts: Ramel, 2008); (3) XML encoding of texts http://www.bvh.xn--univ-tours-tt6e.fr/XML-TEI/index.asp.%20(4)%20%20DissimiLog (dealing with the old usage of i/j and u/v alternation: *vne iambe* morphing to *une jambe*, *viure enyure* morphing to *vivre enivré*); (5) Analog, (Lay, 2010), a tool for lemmatization and morphosyntactic tagging.

B. The normalization of spelling variation

In a virtual environment the reader expects to access the content of the document through functions based on the retrieval of character strings. But this is of very limited use for Renaissance texts as shown above. The character strings aimed at are in fact all those that correspond to the intention of the query: this is what happens when the sought form has only one spelling.

One possibility is to enrich texts with linguistic information gained from lemmatization and morphosyntactic tagging. The forms retrieved, whatever their spelling, are lemmatized under a canonic form which then becomes the pivot of further requests: for example the lemma for *nuit* groups together forms like *nuit* (which is "regular french") or *nuyctz* (old written form).

A first solution, Humanistica (Lay, 2000) was based on the adaptation of a probabilistic tagger/lemmatizer. Several specificities from the VHL project leads to another solution, viz. the development of an environment helpful at all stages of corpus observation, lexical resource creation and text annotation. This tool, called AnaLog (Lay, 2010a, 2010b), is currently being used in the framework of a new "enriched" virtual library.

But the enrichment of text through linguistic annotation is a slow and costly process. Though this solution is very useful to go on producing a reference environment slowly but safely, it is nonetheless desirable to avail oneself of efficient research tools on corpora of texts already available but not yet annotated.

Consequently, in this heterographic environment, our aim is to spot all the written forms which could correspond to a query, being insensitive to variation, without requiring the lemmatization process.

To solve this problem one has to go back to observational evidence, viz. the texts which are the targets of searches. The variability they exhibit must be measured precisely. Two directions may be taken in this respect: either observe the texts or observe the variants attested for a given form.

(1) Concerning text observation, the aim is to evaluate the number of forms for a given text which do not correspond to the norm. Moreover, one must take into consideration the extent to which the texts can be compared. We intend to illustrate this with two short extracts from Montaigne and Rabelais, two authors of paramount significance.

(2) Concerning the observation of variants attested for one word, the idea is to formulate the rules which govern the production of abnormal forms.

II. CHARACTERISATION OF EQUIVALENT FORMS

A. Observational evidence

The observation of texts with spelling variants dating back to a time before the invention or generalization of prescriptive spelling helps perceive how frequent the phenomenon was. Here is an extract from Montaigne, in which the "unexpected" forms (for a French-speaking contemporary reader) are printed in bold:

« De la **coustume** & de ne changer **aisement vne loy receüe**.

Celuy me semble **auoir tres**-bien **conceu** la force de la **coustume**, qui premier forgea ce conte, **qu'vne** femme de village ayant **apris** de caresser & porter entre ses bras **vn** veau **des** l'heure de sa naissance, & continuant **tousiours a** ce faire, **gaigna** cela par **l'accoustumance** que tout grand **beuf** qu'il **estoit**, elle le **portoit** encore. Car c'est a la **verité vne** violente & **traistresse maistresse d'escole**, que la **coustume**. Elle **establit** en nous peu a peu a la **desrobée** le pied de son **authorité**: mais par ce doux & humble commencement l'ayant rassis & planté **auec l'ayde** du temps, elle nous **decouure tantost vn furieux** & **tirannique** visage, contre lequel nous **n'auons** plus la liberté de **haulsser** seulement les yeux. »

Montaigne's text is easily deciphered by the Francophone contemporary reader with no special expertise whereas Rabelais's text is much more difficult to understand. In fact, except the "dissimilation point", the forms extracted from Montaigne's text could be mistaken for spelling errors found in schoolchildren's papers, which is not the case with Rabelais's text:

« Vous **estez deuement adverty**, Prince **tresillustre**, de quants grands **personaiges j'ay esté**, et suis **journellement** stipulé, requis, & importuné pour la continuation des mythologies Pantagruelicques: **alleguans** que plusieurs gens **languoureux**, malades, ou autrement **faschez** & **desolez avoient** a la lecture **d'icelles** trompé leurs **ennuictz**, temps joyeusement passé, & **repceu alaigresse** & consolation nouvelle. »

The conclusion reached after the close study of a substantial set of texts is that the texts themselves comprise a highly unstable environment, hardly compatible with knowledge acquisition strategies based on statistical regularities. Otherwise there are undoubtedly structures which help interpreting the text and which are based on a "regularization" of data and their alignment with contemporary familiar forms. The idea is therefore to observe the lexical items and to detect all the forms they may take, in order to make regular patterns more visible. Here are some of them:

Vices →	Vyces, visces
souverain →	Souverein, souuerain, souverain, soulverain, soulverein, souverayn, soverain, soverein, souvrain, sovrain

If one is to observe the word *souverain* for example, it is possible to identify among the variants encountered a certain number of regular patterns concerning alternations between *i* and *y (vice/vyce), u* and *ul (autre/aultre)* or also *a* and *e* in certain contexts. The identification of such regular patterns and the observation of long lists of examples give the impression of a dense jungle of possible targets, the combination of possible substitutions skyrocketing as the word gets longer.

B. Spotting a word in a given text

In order to figure out a way to move forward, let us keep in mind our initial objective: to formulate requests which would provide all the forms corresponding to the requested word. The next stage is therefore to compare a list of "words to be searched" with their actual occurrences in the text: this is a form of contextualization. Let us go back to the text extracted from Montaigne's work and try to spot the following words:

> *"pied, nature, raisons, reçue, appris, celui, toujours, tantôt, autorité, épée, enivrés, maîtresse, école, établi, allégresse, loi, âge, enfant, pays, mâles, fait, livrer, savons, fouettés, mets, fasse, reine, impératrice, empiéter, prise, vue"*

Comparing the searched forms and their spelling in text, a typology of the situations occurring may be offered. The form being searched is sometimes that which does occur in the text (*raisons/raisons*); in some cases the link seems to be very weak (*impératrice/empériere*), and between these two types a whole gradation of situations can be organised on a linguistic basis.

III. CONCEPTION OF RULES

A. Presentation of the typology

Obvious findings:

- Tokens similar to the type :

pied > pied *nature > nature* *raisons > raisons*

- Well-known phenomena.

(a) Some of these are "transparent" because morphological or derivational traces of them may be found in contemporary French:

(a-1) Based on such derivational sets

It does not come as a surprise that many occurrences of the circumflex ("^") should be equated to an occurrence of *s* :

maîtresse > maistresse *tantôt > tantost*

(a-2) Similarly, there are many occurrences of *é* which appear as *es*. For example,one can find in contemporary spellings : *étude, étudiant, estudiantin, studieux*

école > escole *épée > espée* *établi > estably*

(a-3) It could also be a *c* occurring between a *i* and a *t* : *fruict, faict, dict, effect, nuyct*. In contemporary French, the *c* may occur in a word from the same "family" : *fruit/fructueux ; fait / facture ; dit / dicton ; nuit / nocturne ; effet/effectif*. It is therefore quite easy to "guess" that one can let the *c* away.

(b) Some of these are "transparent", based on homophony :

(b-1) Occurrences of *i* are also interchangeable with those of *y* and vice versa:

celui > celuy *loi > loy* *pays > pais*

(b-2) Different ways of writing the [ɛ] are still possible in contemporary French, they may alternate : règle, *reigle, fait, faict, fouet, foët, prête, preste, alegresse, alaigresse*

(b-3) The same thing follows for [ã] : *ampieter, empiéter, melancolye, mellencolie, mélancolie, semble, samble*

(b-4) Variation also affects consonants and double consonants, with several spellings for the same consonant sound (resulting in what are still considered as common spelling errors):

appris > apris *empiéter > ampieter*

face > fasse // fasse > face *autorité > authorité*

- Regular but obsolete way of spelling

Some regular alternations do not occur any longer. Some cases are more difficult because they are further apart from contemporary production in some way or other, even though they remain legible.

(a) The problem of alternations that are now normalized in modern editions, notably those between *u* and *v* and between *i* and *j*:

livrer > liurer *enivrés > enyures*

(b) Multiple alternations in spelling

Inflections, especially verb inflections, also provide examples of variation: *portais, portay, portois*, etc. are well-known examples but there are other examples such as the past participles written *eu* instead of *u :*

reçue > receüe *vue > veüe* *lu > leu*

(c) Frequently used verbs

Some verbs, frequently used in discourse are really puzzling:

savons > sçauons *prise > prinse* *né > nai*

These verbs are also irregular in contemporary French. And so is a part of the most frequently used vocabulary such as the verbs *avoir, être, aller, faire*.

- Cases for which little can be done

In a certain number of cases already mentioned, the variation is on the borderline between spelling variation and morphological variation. Once the borderline is clearly traced, it may seem legitimate to abandon the hope of identifying two forms: thus *Emperière*, (the feminine of *Empereur)* will not be identified as a variant of *Impératrice.*

B. Formalizing substitution rules

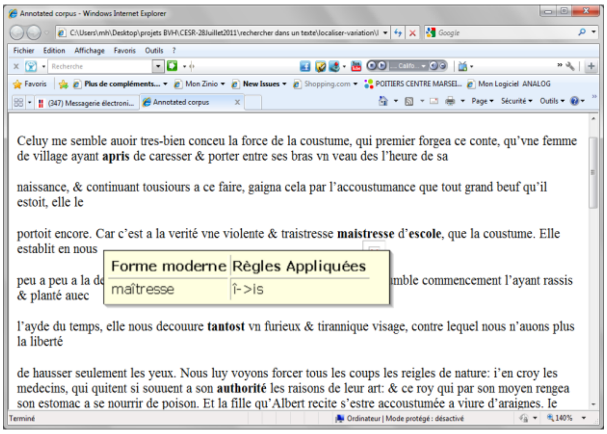
The situation seems to be rather simple "intellectually": due to the structural instability of this linguistic data, equivalences between character strings are difficult to track statistically and no model-based approach can be developed. But linguistic knowledge helps recognize regular replacement patterns, which can be turned into rules. The next point which needs to be taken into account is the relevance of the detected substitution rules: do they really help find all the forms concerned (low silence, good recall), and do they avoid generating too much noise (good precision)? In fact, one may generate a large number of candidate words which are possible in terms of calculation, but totally unacceptable from a linguistic point of view (*autorité <> àüttolrrythêz*). Obviously, there is a lot of noise with no linguistic relevance, and linguistic relevance is an important point in the BVH/VHL context.

The solution chosen to fix that problem is to describe, for each rule, the context in which the substitution is allowed. This aims at constraining their application strongly, and limits their productivity. This contextualisation is based on a good knowledge of the linguistic process involved.

For example, `(?<\=[aeiouy])c(?\=[eiy]) = ss` express that a "*c*" can be changed to "*ss*" when it appears between two vowels and the second one is a palatal vowel. So that the word *face* can be written *fasse* The results now meet our expectations: the rules produce all the linguistically permissible variants.

IV. DESCRIPTION OF THE TOOL

The tool itself is thought to be really user-friendly especially for the tuning of rules and the evaluation of their consequences (efficiency and non regression tests). It is a java program which first transforms a list of words into an extended list of forms, using that for a rules set. Having done this, the need is to localise the different forms attested in the old spelling in a text, according to the requested form. Hence one can identify two "phases". (1) Generating the extended form of the request: at this step, the program generates 3 files; two of them are dedicated to synthetic information about the process (using the rules: how often they have been used) and the end result (how many generated forms). The third one is a file containing, for each word, the list of the generated forms as well as the rules used in the process. This information is really useful to tune the sea of rules: to detect false written rules, conflict between rules, redundancies, noisy ones and so on. (2) Finding the right form within a text: when the extended request is calculated, the ultimate test is to identify all the variants really attested in the text. This is the second phase of our program. The output file of this last part of the process is an html file with a graphical highlighting (or bold character) of the identified variant. Moreover, each form is connected to a bubble showing the rules used to derive the variant. A table containing the summary of the used rules for the text is also available. So, the human validation process is quite friendly.


Annotated corpus - Windows Internet Explorer
C:\Users\mh\Desktop\projets BVH\CESR-28Juillet2011\rechercher dans un texte\localiser-variation\l
Google
Fichier Edition Affichage Favoris Outils ?
Recherche
Favoris
Plus de compléments...
Mon Zinio
New Issues
Shopping.com
POITIERS CENTRE MARSEI...
Mon Logiciel ANALOG
(347) Messagerie électroni...
Annotated corpus
Page
Sécurité
Outils
Terminé
Ordinateur | Mode protégé : désactivé
140%

Celuy me semble auoir tres-bien conceu la force de la coustume, qui premier forgea ce conte, qu'vne femme de village ayant **apris** de caresser & porter entre ses bras vn veau des l'heure de sa

naissance, & continuant tousiours a ce faire, gaigna cela par l'accoustumance que tout grand beuf qu'il estoit, elle le

portoit encore. Car c'est a la verité vne violente & traistresse **maistresse** d'**escole**, que la coustume. Elle establit en nous

peu a peu a la d[…]mble commencement l'ayant rassis & planté auec

Forme moderne	Règles Appliquées
maîtresse	î->is

l'ayde du temps, elle nous decouure **tantost** vn furieux & tirannique visage, contre lequel nous n'auons plus la liberté

de hausser seulement les yeux. Nous luy voyons forcer tous les coups les reigles de nature: i'en croy les medecins, qui quitent si souuent a son **authorité** les raisons de leur art: & ce roy qui par son moyen rengea son estomac a se nourrir de poison. Et la fille qu'Albert recite s'estre accoustumée a viure d'araignes. Ie

For now, with 100% recall, our expectations are fully fulfilled. The testing phase will be continued.

V. CONCLUSIONS

The highly "instable" situation of spelling at the time of the French Renaissance is not conducive to the acquisition of rules by automatic procedures: on the one hand, the evolution caused by the passage of time is significant as the period extends over two centuries which have been marked by mutations in the French language and its standardization; on the other hand, variation is observable on one and the same page, in one and the same book, and may change radically from one book to another or even from one edition to another and each author has a different point of view in this respect.

With a little training, the reader is led to elaborate a certain method combining the different possible pronunciations for a given string of letters. To make it simple, this seems to be done by substituting the one pronunciation for the other on the basis of substitutions otherwise possible in contemporary French. The reader resorts to his knowledge of lexical structures and of derivations in word formation, exploring his mental lexicon and its organisation, and relying heavily on this "engine of mental approximation" which leads us to identify an existing form, occasionally running the risk of making a mistake. The structuring and formalisation of these procedures help establish a powerful system of rules which must be "tamed" to restrict their application to linguistically relevant situations. This yields an extension procedure for requests to be integrated to the tools available on the web site of the BVH/VHL: XTF, PhiloLogic, AnaLog.

REFERENCES

Burnard L. (1995) Text Encoding for Information Interchange – An Introduction to the Text Encoding Initiative, in *Proceedings of the Second Language Engineering Conference*.

Demonet, M.-L. (1996), Pronostiquer avec Hyperbase, *Mots chiffrés et déchiffrés*, Slatkine, pp. 455-471.

Demonet, M.-L., Lay M.-H. (2008), Digitizing European Renaissance prints: a 3-year experiment on image-and-text retrieval, Kolkata, (IWDPH07)

Habert B. & Zweigenbaum (2002), Régler les règles, *TAL*, 43-3, pp. 83-105.

Kraif O. (2011) Les concordances pour l'observation des corpus : utilité, outillage, utilisabilité, in *Actes des 10 ans de la MSHS de Poitiers*, Rennes: PUR.

Lay (Antoni), M.-H., Demonet M.L. (2000) Adaptation d'un lemmatiseur au corpus rabelaisien : naissance d'Humanistica. Jadt 2000

Lay, M.-H., Pincemin B. (2010) Pour une exploration humaniste des textes : AnaLog in jadt 2010

Ramel, J.-Y., Busson, S., Demonet, M.-L. (2006) AGORA: the interactive document image analysis tool of the BVH project, *DIAL*, Digital Image Analysis for Library, Lyon.