Un corpus de textes français pour l'analyse de la variation diachronique et dialectale[1]

France Martineau

Département des lettres françaises, Université d'Ottawa
fmartin@uottawa.ca

ABSTRACT. We discuss two projects dealing with the analysis of French texts from a diachronical and dialectal perspective. The project *Chevalier au Lion* (LFA) presents various manuscripts of *Chevalier au Lion*, with different modules for their analysis (indexes, lexicon, grammatical database). The *Base d'analyse verbale*, with a FileMaker Pro interface for the Web, is an efficient tool for morphologic analysis, but another program, SATO, is used to execute more complex syntactic and contextual tasks. The project *Microvariation et épistolarité en Nouvelle-France* presents a corpus of 17th, 18th and 19th century texts written in vernacular French. The results of automatic lemmatisation programs such as Tree Tagger are poor due to the high spelling variations. The solution considered is a program which recognizes writing strategies used by less educated people.

KEYWORDS : Filemaker Pro, SATO, medieval French, vernacular French

RÉSUMÉ. Notre article présente deux projets sur l'analyse de corpus français dans une perspective diachronique et dialectale. Le projet *Chevalier au Lion*, sur le site du LFA, a pour objet une présentation des différents manuscrits du *Chevalier au Lion*, avec des modules d'analyse (index, lexique, bases grammaticales). La *Base d'analyse verbale* qui a été constituée avec File Maker Pro et une interface Web s'avère un outil comparatif performant pour la morphologie ; toutefois, pour des interrogations contextuelles, SATO demeure un outil plus efficace. Le projet *Microvariation et épistolarité en Nouvelle-France* présente des textes en français familier des XVIIe, XVIIIe et XIXe siècles. Le problème de la lemmatisation automatique est rendu encore plus aigu par la graphie non standard ; nous développons un logiciel intégrant les stratégies d'écriture des illettrés.

MOTS-CLÉS : variation graphique, lemmatisation, FileMaker Pro, SATO, bases d'analyse verbale, français médiéval, français vernaculaire.

1. Constitution de corpus de textes électroniques

Les corpus de textes sont de plus en plus nombreux sur le Web, mais pour le linguiste travaillant sur le français, ces corpus présentent souvent des inconvénients: ils sont limités à une époque du français ou à un français littéraire, et le contexte d'analyse critique fait parfois défaut.

[1] Cette recherche a reçu l'appui financier du CRSHC (410-2001-0119).

Depuis 1996, le *Laboratoire de français ancien*[2] offre aux usagers du Web une collection de textes numérisés (une quinzaine de textes) à partir d'originaux (manuscrits pour l'ancien et le moyen français; premières éditions pour la langue de la Renaissance et le français classique) (ex. : *Couronnement de Louis, Chevalier au Lion, Miracles de Notre-Dame* tirés du *Rosarius, Le Diable boiteux*). Lié à ce site se trouve le site du TFA (*Textes d'ancien français*)[3] , hébergé à l'ARTFL, qui comprend plusieurs textes numérisés pouvant être interrogés avec le moteur de recherche de l'ARTFL. Contrairement à l'ARTFL (ou au TFA), l'objectif du LFA n'est pas tant la masse critique de textes, dont la valeur n'est pas remise en question pour les analyses statistiques d'envergure sur des stades du français, mais la présentation de textes avec développement de leur apparat analytique : manuscrits originaux (avec les photos numérisées du manuscrit), éditions critiques, index et lexiques, bases de données, articles. Le projet du *Chevalier au Lion*, que nous présentons en Section 2, a été conçu dans cet esprit.

Dans la même perspective, nous avons créé un autre site, *Microvariation linguistique et épistolarité en Nouvelle-France*[4] , qui prend la suite chronologique du LFA. Ce site couvre la période des XVII^{e}, XVIII^{e} et XIX^{e} siècles. Il est unique en ce qu'il propose des lettres inédites écrites en français familier, avec une répartition diachronique (du XVII^{e} siècle au début du XX^{e} siècle) et dialectale (Nouvelle-France, Acadie, Nord-Ouest de la France) alors qu'un site comme celui de l'ARTFL se fonde essentiellement sur des textes littéraires. Nous présentons ce site en Section 3.

Pour chacun des deux projets, nous discutons brièvement de l'édition même du texte puis nous examinons les solutions et les problèmes posés par l'intégration des autres modules autour du pivot que constitue le texte.

2. Le projet du *Chevalier au Lion*

Le *Chevalier au Lion* est un des chefs-d'œuvre de Chrétien de Troyes, un des auteurs les plus importants de la période médiévale. Le projet *Chevalier au Lion* s'inspirait, à l'origine, du projet *Charrette* [5] de l'Université Princeton portant sur un autre roman de Chrétien de Troyes. Il a paru intéressant de comparer deux textes d'un même auteur, pour lesquels plusieurs manuscrits ont eu le même copiste.

Dans le projet *Chevalier au Lion*[6], dirigé par Pierre Kunstmann et moi-même, nous avons essayé de créer un va-et-vient entre les manuscrits d'un même texte, l'édition critique, les index, les lexiques et les bases d'analyse grammaticale avec outil d'interrogation sur des aspects de la langue (variantes dialectales, valence du verbe, discours rapporté).

La première étape a été de fournir à l'usager le texte lui-même, dans ses différentes versions manuscrites. Nous disposions de 9 manuscrits et de fragments ; pour l'instant 7 manuscrits ont été transcrits[7] . Le texte est présenté avec l'image des feuillets pour deux manuscrits, le manuscrit H et celui de Princeton.

---

[2] http://www.uottawa.ca/academic/arts/lfa
[3] http://www.lib.uchicago.edu./efts/ARTFL/projects/TLA/
[4] http://www.uottawa.ca/academic/arts/nf
[5] http://www.princeton.edu/~lancelot/
[6] http://www.uottawa.ca/academic/arts/lfa/activites/textes/chevalier-au-lion/chlpresduprojet.html
[7] http://www.uottawa.ca/academic/arts/lfa/activites/textes/chevalier-au-lion/index.html

* Fr. 794 (H) - Avec images des feuillets.
*Fr. 1433 (P).
*Fr. 1450 (F).
*Fr. 1638 (L).
*Fr. 12560 (G).
*Fr. 12603 (S).
*Chantilly, Musée Condé 472 (A).
*Vatican, Christine 1725 (V).
* Princeton, Garrett 125 (R) - Avec images des feuillets.

Le va-et-vient entre les différents manuscrits nécessitait aussi qu'un manuscrit serve de base, à partir de laquelle les différences pouvaient être établies. Le manuscrit H, celui du copiste Guiot, a été choisi parce qu'il est considéré comme l'un des meilleurs par les spécialistes de Chrétien de Troyes [WOLEDGE 1986, t. I, p. 2], aussi parce que la langue champenoise du copiste Guiot se rapproche beaucoup de celle de l'auteur Chrétien de Troyes. Les vers de la transcription ont été numérotés, de façon à servir de référence de base.

On peut ainsi lire le texte, vers après vers, de façon continue[8]. Le texte est divisé selon les 'histoires' ou les 'épreuves' du *Chevalier au Lion*, de façon à limiter le temps de téléchargement. Le début du manuscrit H apparaît de la façon suivante :

Manuscrit H 79d.
1. Artus, li boens rois de Bretaingne,
2. La cui proesce nos enseigne
3. Que nos soiens preu et cortois,
4. Tint cort si riche come rois
5. A cele feste qui tant coste,
6. Qu'an doit clamer la Pantecoste.

Comme la transcription de chaque manuscrit présente un traitement isolé, l'utilisateur intéressé aux variantes entre les manuscrits peut trouver le travail de comparaison fastidieux. Nous avons donc fait équipe avec Kajsa Meyer, de l'Université de Copenhague, qui a indiqué, vers après vers, les différentes versions selon les manuscrits[9]. Ci-dessous, le premier vers du texte :

[8] http://www.uottawa.ca/academic/arts/lfa/activites/textes/chevalier-au-lion/H/Hpresentation.html

[9] http://www.uottawa.ca/academic/arts/lfa/activites/textes/kmeyer/kpres.html

v.1-200
H 1 ****A**rtus li boens rois de Bretaingne /79v°a/
P 1 ****L**i boins roys Artus de Bretaigne /61r°a/
V 1 ****L**i bons rois Artus de Bretaigne /34v°a/
F 1 ****L**i bons rois Artus de Bretaigne /207v°b/
G 1 ****A**rtus li bons rois de Breteigne /1r°b/
A 1 ****A**rtus li boins rois de Bretaingne /174r°a/
S 1 ****A**rtus li boins rois de Bretagne /72r°a/
R 1 ****A**rtus li boins rois de Bretagne /40r°a/
Ly 1 ****L**i bo..s rois Artus de Bretaigne /1r°/

Dans son état actuel, ce projet ne comprend pas d'outils d'analyse grammaticale ; il offre tout au plus des outils pour une analyse comparative, qui devra encore faire beaucoup appel au balisage manuel du texte. P. Kunstmann a donc créé un index lemmatisé de base pour le manuscrit H[10], comportant le lemme, l'indice grammatical (verbe, nom, préposition, etc.), les formes occurrentes, la fréquence et la référence. Il s'est servi à la fois des travaux de Marie-Louise Ollier (1986) sur ce même manuscrit et du logiciel de concordance *Concord Oxford*. La lemmatisation est essentiellement manuelle, le logiciel servant à classer les formes. Ci-dessous, une partie de la forme cuidier 'penser' :

CUIDIER v cuida (3) 1956, 3383, 5451 cuidai (7) 323, 444, 478, 1268, 3117, 3649, 6237 cuide (14) 882, 1162, 1751, 2195, 2587, 3030, 3086, 3498, 4223, 4853, 5426, 5816, 5843, 5853 cuident (2) 2460,
5347 cuiderent (1) 5348 cuideroie (1) 5156 cuideroit (1) 6590 cuideront (1) 1067 c uidier(5) 87, 533, 1428, 3054, 5071 cuidiez (7) 75, 1602, 1676, 1698, 2437, 5488, 6523 cuidoient (1) 1093 cuidoit (8) 677, 2727, 2743, 4100, 4755, 6655, 6657, 6679 cuit (34) 77, 95, 212, 335, 414, 472, 997, 1068, 1127,

Cet index, utile pour ceux qui s'intéressent au lexique, demeure malgré tout assez statique pour l'analyse grammaticale et ne met pas en évidence la variation entre les différents manuscrits du *Chevalier au Lion*. C'est pourquoi nous avons voulu constituer une base d'analyse verbale qui aurait les avantages à la fois d'un index lemmatisé et de la présentation de K. Meyer. Il fallait donc un logiciel qui permette de faire :

i. une analyse morpho-syntaxique des principaux verbes et de leurs arguments;
ii. une étude comparative des verbes dans les différents manuscrits du *Chevalier au Lion*;
iii. une étude de l'ordre des mots.

Pour atteindre ces objectifs, nous avons utilisé deux logiciels aux fonctions différentes : FileMakerPro avec une interface Web et SATO Web.

[10] http://www.uottawa.ca/academic/arts/lfa/activites/travaux_ling/chlindex/Chlindexlem.html

2.1 FILEMAKER PRO: Pour une analyse morpho-syntaxique de base

Base d'analyse verbale du Chevalier au Lion[11]

Le logiciel FileMakerPro, doté d'une interface Web conviviale, a servi à créer des fiches. Le choix des verbes a été guidé par leur classe sémantique (par exemple, les verbes d'opinion ou de volonté). À chaque occurrence d'un verbe correspond une fiche présentant des variables morpho-syntaxiques : graphie, lemme, temps, mode, personne, voix, construction personnelle/impersonnelle/inaliénable. La référence dans l'entrée de chaque fiche peut ramener l'utilisateur au vers de la transcription, ménageant ainsi un lien entre le module *Base d'analyse verbale* et celui *Texte du manuscrit*.
On peut faire des interrogations simples:

Chercher le lemme 'cuidier'
Chercher toutes les formes à l'imparfait

ou des interrogations avec plus d'une variable:

Chercher le lemme 'cuidier', personne 3, indicatif

Ce type d'interrogation est particulièrement utile pour l'analyse des graphies verbales et pour la lemmatisation. On peut ainsi remarquer que le manuscrit H de Guiot présente une très grande stabilité des formes graphiques, contrairement au manuscrit P qui affiche une plus grande variation.

FileMaker Pro permet aussi d'ajouter sur une même fiche un nombre important de critères, ce qui a permis d'intégrer une analyse comparative du comportement des verbes, d'un manuscrit à l'autre. Comme le manuscrit H servait de manuscrit de base, un renvoi à celui-ci a été ajouté à chaque fiche, peu importe la source du manuscrit.

Il est ainsi possible de comparer les variantes graphiques, morphologiques, syntaxiques, lexicales ou stylistiques dans les trois manuscrits examinés jusqu'à maintenant (manuscrits H, V et P). À titre d'exemple, on peut examiner la question de trois synonymes en ancien français : *cuider*, *croire* et *penser*.[12] Lavis (1973) notait déjà la prédominance de *cuidier* sur les deux autres verbes et une recherche rapide dans les manuscrits confirme cette tendance.

Notre *Base d'analyse verbale* permet de pousser un peu plus loin l'analyse en comparant le choix du verbe, pour un même contexte, dans les manuscrits du *Chevalier au Lion*. La recherche : *Cherche = Cuidier* sous l'entrée *Lemme* dans le manuscrit H, (*Texte H*), donne accès à toutes les occurrences de ce verbe et son équivalent dans les autres manuscrits. On pourra répéter la recherche pour *penser* ou *croire*.

On voit d'abord que les manuscrits H et P présentent très peu de variation entre eux du point de vue des variantes entre les trois verbes; là où le manuscrit H affiche *cuidier*, le manuscrit P affiche le plus souvent le même verbe. Il n'est pas étonnant que leurs similitudes soient grandes (bien que le manuscrit P soit fortement teinté de picard) puisqu'ils appartiennent à la même branche et ont probablement comme point de départ un même manuscrit [WOLEDGE 86].

---

[11] http://www.citemax.net/uottawa/france1/index.html
[12] Une étude plus approfondie de l'emploi de ces trois verbes dépasse les objectifs de cet article.

C'est donc entre les manuscrits H et V que l'on note des écarts. Ces deux manuscrits appartiennent à deux branches différentes, et le copiste du manuscrit V s'écarte souvent du manuscrit H par une volonté d'alléger le texte [WOLEDGE 86, citant JONIN]. La relation entre les deux manuscrits est aussi mise en évidence dans l'emploi des trois verbes synonymes. Notons d'abord que *cuidier* dans le manuscrit H alterne avec *croire*, et non avec *penser*, dans le manuscrit V. En fait, *penser* dans le manuscrit H n'alterne pas avec un autre verbe dans le manuscrit V, ce qui semble conforter l'hypothèse que *penser* ne fait pas partie de la relation *cuidier/croire*. D'autre part, le copiste du manuscrit V semble donner un sens particulier à croire puisque, s'il apparaît déjà dans le manuscrit H, il ne le change pas dans le manuscrit V. En d'autres mots, pour le copiste du manuscrit V, *cuidier* semble être un équivalent possible de *croire*, mais l'inverse n'est pas vrai. Le tableau qui suit résume la relation entre les trois verbes :

Manuscrit H	Manuscrit V	Manuscrit P
Cuidier	Cuidier/Croire	Cuidier
Croire	Croire	Croire
Penser	Penser	Penser

Certains éléments syntaxiques ont également été incorporés dans la fiche, soit l'analyse de la valence (réalisée ou maximale), la présence d'autres compléments, la présence d'un pronom de reprise et l'ordre des mots.

Dans le cas de la valence verbale, le nombre d'arguments, qu'ils soient potentiels ou réalisés, comme dans l'exemple suivant, a pu être intégré sans problème:

***Valence maximale** : 2*
***Valence réalisée** : 1*
***Citation** :*
*Ne **cuit** qu'an plain ne an boschage*
Puisse an garder beste sauvage,
N'en autre leu, por nule chose,

Toutefois, nous nous sommes heurtés à un problème de nature contextuelle. Comment savoir, pour un verbe qui a une valence maximale 2 et qui ne réalise qu'un seul de ses arguments, lequel des arguments (sujet ou verbe) est réalisé ? Les informations contextuelles, comme la fonction des arguments du verbe et leur ordre dans la phrase, sont gérées avec difficulté par FileMaker Pro. Pour récupérer dans une certaine mesure cette information, nous avons ajouté, en plus de la citation, une entrée dans laquelle est indiqué l'ordre des mots des constituants majeurs, comme le montre l'exemple suivant :

***Ordre des mots** : (Sujet) V COD*

Cette entrée est traitée comme du texte, et les recherches du type *Cherche tous les verbes avec sujet effacé et suivi d'un COD* fonctionnent en autant que la suite linéaire

corresponde exactement à l'interrogation (sans élément intervenant, comme un adverbe par exemple).

Il aurait été possible de créer une fiche pour permettre une interrogation contextuelle plus précise mais seulement au prix d'une complexification de la fiche. C'est pourquoi nous nous sommes tournés vers le logiciel SATO, conçu par François Daoust, au Centre ATO de l'Université du Québec à Montréal, dirigé par Monique Lemieux.

2.2 SATO Web : Pour une analyse contextuelle[13]

Le logiciel **SATO** permet des interrogations contextuelles de façon plus aisée. Daniel Labonia et l'auteure préparent une *Base d'analyse verbale du Chevalier au Lion sur SATO* qui reprendra les traits morpho-syntaxiques définis dans la *Base d'analyse verbale FileMakerPro*[14]. Le lexique établi sera projeté sur le texte pour coder les arguments du verbe (leur expression, leur valeur et leur position par rapport au verbe).

Ainsi, l'occurrence *cuit* du lemme *cuidier* apparaissant dans le *Texte* est définie dans le *Lexique* à partir de l'information récupérée dans la *Base d'analyse verbale FileMakerPro* de la façon suivante :

Texte *:*

Ne ***cuit*** *qu'an plain ne an boschage*
Puisse an garder beste sauvage,
N'en autre leu, por nule chose,

Lexique :
Cuit : lemme *cuidier*, personne 1, temps présent, mode indicatif, voix active, construction personnelle.

L'information contextuelle est ensuite directement indexée sur le mot *cuit* dans le texte :

Information contextuelle indexée **:**
Verbe: valence 2
Sujet: non-exprimé
Complément: exprimé, subjonctif, postverbal

L'interaction des trois modules permet d'interroger le texte sur l'ordre des constituants, soit à partir de leur fonction ou de leur nature grammaticale :

Cherche le verbe cuidier suivi d'un complément

Cherche un verbe sans sujet exprimé

Cherche la suite Adv- Verbe

---

[13] http://www.ling.uqam.ca/ato/
[14] Cette recherche devrait être bientôt disponible sur le site du LFA.

3. Un corpus de français québécois familier du XVII^{e} siècle au début du XX^{e} siècle

Les textes du LFA couvrent essentiellement la période de l'ancien et du moyen français. Les études diachroniques sur le français à des époques postérieures doivent se fonder sur des corpus en majorité composés de textes littéraires dont la langue a subi la standardisation associée à leur époque, ce qui pourrait être en partie la cause des changements linguistiques si brusques notés dans les études linguistiques pré- et post-XV^{e} siècle. Le projet *Microvariation et épistolarité en Nouvelle-France* permet de mettre en lumière la présence de micro-systèmes associés à des registres différents de langue et donne un tableau plus juste de l'évolution diachronique du français.

Le corpus est constitué de correspondance écrite en français familier (entre membres d'une même famille, par des soldats peu éduqués, etc.), au Québec, en Acadie et dans les régions du Nord-Ouest de la France durant les XVII^{e}, XVIII^{e} et XIX^{e} siècles. Ce type de corpus est, à notre connaissance, inédit. S'il existe des corpus de textes littéraires en français familier [AYRES-BENNETT 00 ; LODGE 02), les corpus de français de gens peu éduqués sont rares, et lorsqu'ils existent ils se limitent à un journal personnel d'une seule personne (voir ERNST 02).

Ce corpus permet de suivre diachroniquement les variations dans un français populaire ou familier et de comparer les variantes dialectales pour une époque donnée. Ces textes de français familier offrent certes une reconstitution partielle du français parlé de l'époque [BRANCA-ROSOFF 94 ; WÜEST 85), mais comme les formes parlées sont perdues à jamais, ces textes présentent de précieux éléments pour notre compréhension de l'évolution du français.

3.1 Constitution et saisie électronique

La constitution du corpus lui-même a pris plus de 10 ans de recherche dans les archives du Québec, de l'Acadie et de la France à la poursuite de textes écrits en français familier. La saisie électronique de plus de 500 lettres en français vernaculaire, souvent difficiles à déchiffrer parce que écrites par des illettrés, est terminée. Pour l'instant deux textes ont été mis en ligne. L'édition sur le Web comprend l'image du feuillet, la transcription de l'original et sa traduction moderne[15]. À titre d'exemple, voici le texte acadien :

[15] http://www.uottawa.ca/academic/arts/lettres/nf/Base%20textuelle.htm

Transcription de l'original	Transcription moderne
Cher pere Cet a present que ge pran le plesir de vous écrire de mes nouvelle que ge sui an bone sante dieu marci et g espère que vous lette aussi bien que moi E tout lecipage Le temps et bien mauvait, il vante tout les deux ou trois giour tro pour pecher quand nous on arives la moru etait bien rare mais ge comman- son a prendre courage pour notre charge nous avon apresent 130 quento & grand mesure tout les batiman de par che nous Son comme nous otre Nous avon Eeu la malle chance augiourdhui de pardre une ancre. Et setout la varie que nous avon Eeu De plus nos petit canos sont bien vallan Ge fini an vous fesan mes compliment a toute la famille Silvin mande au vieux quil Voudrai bien qui eu la bonte De envoyer des nouvelle a madelene Quil et bien vallan	Cher père c'est à présent que je prends le plaisir de vous écrire de mes nouvelles que je suis en bonne santé dieu merci et j'espère que vous l'êtes aussi bien que moi Et tout l'équipage Le temps est bien mauvais, il vente tous les deux ou trois jours trop pour pêcher quand nous on arrive la morue etait bien rare mais je commençons à prendre courage pour notre charge nous avons à présent 130 quento & grand mesure tous les bâtiments de par chez nous Sont comme nous autres Nous avons eu la malchance aujourd'hui de perdre une ancre. Et c'est tout l'avarie que nous avons eu De plus nos petit canots sont bien vaillants Je finis en vous faisant mes compliments à toute la famille Silvain mande au vieux qu'il Voudrait bien qu'il eut la bonté De envoyer des nouvelles a Madeleine Qu'il est bien vaillant

3.2 Lemmatisation

Pour la lemmatisation des textes d'ancien et de moyen français, nous pouvions compter sur des dictionnaires et sur une certaine stantardisation des variantes selon la région d'origine des manuscrits. Les textes d'écrivants peu éduqués présentent aussi de nombreuses variantes orthographiques, mais les auteurs de ces lettres ne suivent pas un système de variation connu.

Des essais ont été effectués avec un logiciel de semi-lemmatisation automatique (Tree Tagger) conçu par Achim Stein de l'Université de Stuggart (achim.stein@po.uni-stuttgart.de). Le résultat de la lemmatisation est décevant : pas plus de 32,4% de réussite sur la catégorisation grammaticale. Une partie des résultats apparaît dans le tableau qui suit :

Cher	ADV	cher
pere	NOM:masc:sg	<unknown>
Cet	DET:DEM:masc:sg	ce
a	VER:sg	avoir
present	VER:pl	<unknown>
que	CON:sub	que
ge	VER:pl	<unknown>
pran	ADJ:masc:sg	<unknown>
le	DET:masc:sg	le
plesir	VER:infi	<unknown>
de	PRP	de
vous	PRO:pl	vous
écrire	VER:infi	écrire
de	PRP	de
mes	DET:femi:pl	mon
nouvelle	ADJ:femi:sg	nouveau
que	CON:sub	que
ge	VER:infi	<unknown>
sui	VER:pper:masc:sg	<unknown>
an	NOM:masc:sg	an
bone	ADJ:femi:pl	<unknown>
sante	VER:pper:masc:sg	<unknown>
dieu	NOM:masc:sg	dieu
marci	NOM:masc:sg	<unknown>
et	CON	et
g	VER:pl	<unknown>
espère	VER:sg	espérer
que	PRO:REL	que
vous	PRO:pl	vous
lette	ADJ:femi:sg	lette
aussi	ADV	aussi
bien	ADV	bien
que	CON:sub	que
moi	PRO:sg	moi
E	NOM:masc:sg	<unknown>
tout	PRO	tout
lecipage	VER:sg	<unknown>

Le	DET:masc:sg	le
temps	NOM:masc:sg	temps
et	CON	et
bien	ADV	bien
mauvait	ADJ:femi:pl	<unknown>
,	PUN	,
il	PRO	il
vante	VER:sg	vanter
tout	ADV	tout
les	DET:masc:pl	le
deux	NUM	deux
ou	CON	ou
trois	NUM	trois
giour	NOM:masc:sg	<unknown>
tro	NOM:femi:sg	<unknown>
pour	PRP	pour
pecher	VER:infi	<unknown>
quand	CON:sub	quand
nous	PRO:pl	nous
on	PRO	on
arives	ADJ:femi:pl	<unknown>
la	DET:femi:sg	le
moru	NOM:femi:sg	<unknown>
etait	VER:sg	<unknown>
bien	ADV	bien
rare	ADJ:masc:sg	rare
mais	CON	mais
ge	ADJ:masc:sg	<unknown>
commanson	NOM:masc:sg	<unknown>
a	VER:sg	avoir
prendre	VER:infi	prendre
courage	NOM:masc:sg	courage
pour	PRP	pour

Il serait possible d'augmenter les chances de réussite du logiciel en y intégrant certaines règles pour rendre compte des régularités que l'on trouve dans un texte écrit en orthographe non standard.

D'abord la non-utilisation des accents dans ce texte pose un problème simple à résoudre. Comme l'accentuation en français sert seulement à distinguer un petit nombre de mots (la troisième personne du passé simple et du subjonctif imparfait, par exemple) [CATACH 1980], il s'agit d'indiquer au logiciel les variantes possibles avec ou sans accent, de façon systématique pour l'ensemble des mots : *etait*, sans accent, sera alors reconnu comme le verbe *être* dans le lexique.

Certaines règles phonétiques bien connues causent aussi des problèmes d'identification et pourraient être intégrées à peu de frais : le 'a' de *marci* ou *pardre* renvoie à l'ouverture de la voyelle /e/ suivie de /r/. Le verbe *être*, troisième personne de l'indicatif présent, pourrait avoir été orthographié *et* pour marquer la fermeture de la voyelle. Notons toutefois que *et* se confond alors avec la conjonction de coordination *et* dans le texte, parfois aussi simplement écrite *e*.

Les consonnes et voyelles muettes présentent aussi un cas assez simple à résoudre : il s'agit de dresser la liste des mots contenant une voyelle ou une consonne muette et d'établir les variantes possibles : le mot *trop*, orthographié *tro* dans le texte, est de ce type, tout comme *che* pour *che* et *moru* pour *morue*.

Certains accords morphologiques peuvent être considérés de la même façon que les consonnes muettes. C'est le cas du pluriel dans le syntagme nominal. Dans *nos petit canos*, il faudra prévoir les séries de variantes suivantes : no/nos; peti/petis/petit/petits, canot/canots/cano/canots.

La morphologie particulière de la première personne de l'indicatif en acadien devra être intégrée, et une désambiguïsation contextuelle sera nécessaire : *ge commanson* (pour *je commence*) qui se confond avec la première personne du pluriel (*nous avon*) dans ce texte.

Il est possible de gérer ces variantes orthographiques dans un logiciel de lemmatisation en autant qu'on puisse formuler des règles. En ce sens, l'agglutination pose de plus sérieux problèmes à la lemmatisation puisqu'elle est assez souvent imprévisible, bien que son contexte d'apparition dans la formation de nouveaux mots ait été bien étudié par les lexicologues (voir entre autres Niederehe 1996 pour les analyses morphologiques non-étymologiques en mitchif ou Brasseur 1996 pour le français de Terre-Neuve)[16]. L'agglutination de l'article dans *lecipage* ou du pronom personnel dans *Lette*, du pronom démonstratif dans *cet* ou *setout* nécessitera une manipulation manuelle pour identifier la division de mots.

Le phénomène inverse, une division de mots non standard, est aussi possible et tout aussi imprévisible : c'est le cas de *malle chance* ('malchance') ou de *la varie* ('l'avarie').

[16] Dans le texte acadien, le rapport entre l'agglutination orthographique et l'analyse morpho-lexicale qu'en fait l'écrivant nécessiterait une recherche plus poussée.

Enfin, d'autres variantes ne semblent obéir à aucune logique graphique. De ce nombre, certaines pourraient sans doute être récupérées par une règle sur les stratégies d'écriture des illettrés ou les habitudes d'enseignement d'écriture ou de lecture de l'époque. Des études ont montré que les enfants en contexte d'apprentissage d'écriture ou de lecture développent des stratégies cognitives face à l'inconnu de l'alphabétisation, des 'écritures inventées' [FIJALKOW 93]. On peut supposer que les écrivants peu éduqués ont aussi recours à des processus similaires. Ainsi, le 's', que l'on trouve dans *on arives*, est peut-être perçu comme un signe de pluriel. La disposition même du texte sur la page, qui peut être analysée grâce à l'image du texte qui accompagne les transcriptions, pourrait aussi permettre d'identifier certaines stratégies dans la séparation des mots, l'emploi des majuscules ou l'accord à distance.

Nous pensons que la lemmatisation d'un texte comme le texte acadien examiné pourra être améliorée par des règles de variation orthographique et que l'analyse de ces règles développera notre connaissance des stratégies d'écriture des écrivants à date ancienne. Ainsi, comment, du point de vue historique et dialectal, les stratégies d'écriture ont-elles varié? Comment ont-elles été influencées par l'enseignement des normes orthographiques de l'époque? À titre illustratif, on peut comparer le texte acadien, écrit en 1869, avec le début d'un texte de la Nouvelle-France, écrit en 1750, qui apparaît aussi sur le site :

a montréal ce 6 juilliette 1750.
mon cher frere Je vien de resevoire votre lestre
du 31 mez Je nan né hu ocune de vous
depuis le 21 7tenbre Lané derniere il fau que
les mien est hu le même sor

Comme dans le texte acadien, l'agglutination est très présente (*nan*, *né*, *Lané*), et l'écrivante omet aussi certaines consonnes muettes (*vien*, *lané*, *fau*, *mien*, *sor*). Mais on note aussi des différences : par exemple, là ou l'écrivant acadien écrit *Eeu* pour le participe passé de *avoir*, l'écrivante de la Nouvelle-France écrit *hu*, signalant ainsi l'origine latine du verbe (*habere*). La lemmatisation devra tenir compte de ces différences par l'élaboration de lexiques distincts dont seulement certaines règles d'écriture seront communes.

Pour l'instant, nous avons indexé les verbes dans le texte au moyen de *DreamWeaver*. Mais le texte, même indexé, demeure assez statique pour l'interrogation. Dans l'esprit de *SATO Web*, nous développons un logiciel qui sera en mesure d'interroger les textes de façon contextuelle et comparative.

4. Conclusion

La présentation de textes sur le Web a pris un essor foudroyant depuis plusieurs années. Il reste toutefois à développer des outils qui puissent mettre en relation les différents modules grammaticaux et lexicaux qui gravitent autour du texte. Un des problèmes auxquels nous nous sommes heurtés avec le LFA et la *Base d'analyse verbale* est la capacité du logiciel à rendre compte du contexte, entendu comme la relation entre des mots dans une phrase ou d'un manuscrit à l'autre.

Le chercheur travaillant sur des langues moins standardisées que le français moderne doit aussi pouvoir résoudre le problème des variantes graphiques, à la base de tout exercice de lemmatisation et d'interrogations grammaticales plus complexes. Le Projet *Microvariation et épistolarité en Nouvelle-France* est en voie de développer un logiciel qui intègrera les récurrences d'écriture des illettrés. Ce logiciel, en plus de fournir un outil précieux à la lemmatisation de textes déviants orthographiquement, permettra de saisir les principes cognitifs d'écriture des illettrés à date ancienne.

Références

[AYRES-BENNET 00] AYRES-BENNETT, W., Voices from the past. *Romanische Forshungen*. 323-348, 2000

[BRANCA-ROSOFF 94] BRANCA-ROSOFF, S. et N. Schneider, *L'Écriture des citoyens*. Paris :Klincksiek, 1994

[BRASSEUR 96] BRASSEUR, Patrice, Changements vocaliques initiaux dans le français de Terre-Neuve. In Lavoie, T. editor, *Français du Canada-Français de France*, Tübingen, Niemeyer, 295-305, 1996

[BURIDANT 00] BURIDANT, C. *Grammaire nouvelle de l'ancien français*. Paris: Sedes, 2000

[CATACH 80] CATACH, N. *L'Orthographe française. Traité théorique et pratique*. Paris, Nathan, 1980

[ERNST 02] ERNST, G. et B. WOLF *Journal de Chavatte*. Niemeyer, 2002

[FIJALKOW 96] FIJALKOW, J. *L'Entrée dans l'écrit*. Toulouse, PUM, 1996

[FURET 77] FURET, F. and J. OZOUF. *Lire et écrire. L'Alphabétisation des Français de Calvin à Jules Ferry*. Paris : Les Éditions de Minuit, 1977

[LAVIS 73] LAVIS, G. La Concurrence entre penser, cuidier et croire chez Chrétien de Troyes. *Marche Romane*, 147-168, 1973

[LODGE 02] LODGE, A. Textes numérisés. (Oxford Text Archive : http://ota.ahds.ac.uk) 2002

[NIEDEREHE 96] NIEDEREHE, H.-J. Le vocabulaire d'origine française du Turtle Mountain Chippewa Cree (Mitchif). In Lavoie, T. editor, *Français du Canada-Français de France*, Tübingen, Niemeyer, 376-386, 1996

[OLLIER 86] OLLIER, Marie-Louise. *Lexique et concordance de Chrétien de Troyes*, Montréal-Paris, 1986

[WOLEDGE 86] WOLEDGE, B. *Commentaire sur Yvain (Le Chevalier au Lion) de Chrétien de Troyes*. Genève Droz, 1986

[WÜEST 85] WÜEST, J. Le patois de Paris et l'histoire du français, *Vox Romanica* 44, 234-258, 1985