*In* Mathieu VALETTE (éd.)
Textes, documents numériques, corpus. Pour une science des textes instrumentée
*Syntaxe & Sémantique*, 9, pp. 87-110.

# Usages linguistiques de la textométrie. Analyse qualitative de la consultation de la Base de Français Médiéval via le logiciel Weblex.

Bénédicte Pincemin[1,3], Céline Guillot[2,3], Serge Heiden[2,3], Alexei Lavrentiev[2,3], Christiane Marchello-Nizia[2,3]

[1] CNRS ; [2] ENS Lettres Sciences humaines ;
[3] ICAR UMR 5191 CNRS & Université de Lyon

Résumé

La textométrie propose des techniques statistiques pour l'étude des occurrences d'un motif linguistique dans un corpus. L'article présente une synthèse des pratiques de linguistes du français médiéval qui ont eu accès à des fonctionnalités textométriques, avec une interprétation au plan linguistique des calculs et formalismes textométriques utilisés. L'enjeu est d'expliquer l'attrait ou au contraire le désintérêt pour chaque fonctionnalité, et d'en tirer des enseignements théoriques et pratiques.

*Abstract*

*Textual statistics offer powerful means to study the realizations of a linguistic pattern in a corpus. This article presents a review of the way linguists studying medieval French through the BFM textual database use textometric software. It provides an interpretation of the textometric calculations based on the linguistic nature of the data. The aim is to understand why some calculations are relevant, and why some others do not interest the linguists, and to draw theorical and practical lessons from this analysis.*

## 1. Problématique

La linguistique a maintenant la possibilité d'observer la langue dans des corpus textuels numériques, dotés de fonctionnalités de consultation systématiques et rapides. Ainsi, en France, la base Frantext[1] est interrogeable par le logiciel Stella, qui permet de relever tous les contextes d'apparition d'un mot, ou d'un motif défini de façon complexe (suite de mots, variations). La textométrie développe les possibilités de consultation et d'analyse de corpus textuels en faisant appel à des décomptes et des modélisations statistiques et en combinant aux possibilités de repérage d'occurrences des calculs de tri, de sélection et de réorganisation statistique. Ces méthodes sont mobilisables pour toute la gamme des paliers de description linguistique (tels que mot, phrase, texte), tant pour les motifs étudiés que pour les contextes mobilisés.

[1] http://www.atilf.fr/atilf/produits/frantext.htm

La Base de Français Médiéval (BFM) propose une édition numérique de 80 textes intégraux (près de trois millions d'occurrences-mots), d'ancien et de moyen français, du IXe à la fin du XVe siècle, avec l'objectif d'offrir au linguiste un corpus représentatif d'usages de la langue écrite (variété des domaines et genres textuels, forme des textes -en vers ou en prose-, diversité géographique). Depuis 1999, la consultation et l'interrogation de la BFM sont proposées via le logiciel de textométrie Weblex.

L'objectif de cet article est de faire un bilan des usages effectifs, par les linguistes, d'un outil de textométrie : proposer une analyse qualitative fondée sur l'expérience acquise par plusieurs années d'exploitation en ligne, interpréter les faits observés tant au plan linguistique qu'au plan des techniques statistiques, en tirer des enseignements qui puissent orienter les développements logiciels et les pratiques. D'une part, il s'agit de mieux comprendre comment le fonctionnement de la langue transparaît et devient saisissable dans le contexte des corpus numériques. On vise ainsi à expliquer la pertinence linguistique des fonctionnalités les plus appréciées, en tenant compte notamment des stratégies et des modalités d'interrogation qui viennent à s'imposer en pratique. D'autre part, nous voulons analyser les écarts entre les attentes des linguistes et les possibilités offertes par la textométrie. Sont-ils à comprendre en termes de manques -manque de fonctionnalités pour l'outil, ou méconnaissance de certaines techniques par les linguistes ? La correction de ces deux types de manques ouvre alors des perspectives essentielles pour de nouveaux développements en linguistique de corpus. Mais dans certains cas, peut-être y a-t-il une inadéquation de fond entre les techniques de la statistique textuelle et certaines approches linguistes : identifier ces décalages contribue à révéler l'essence propre de ces disciplines qui se rejoignent sans se confondre.

Bien entendu, notre enquête est circonscrite à un domaine limité de linguistique de corpus, et un certains nombre d'observations sont contingentes à l'état du développement de la base BFM et de l'outil Weblex[2]. Notre motivation est néanmoins de mettre en évidence des « lignes de force », par convergence d'indices et par une réflexion généralisante ; nous nous attacherons donc à déceler, par delà la singularité de notre contexte, des aspects fondamentaux d'une approche pleinement linguistique des corpus textuels outillés.

## 2. Contexte de l'étude

### *2.1. Le type d'études linguistiques menées sur la BFM*

Nous nous donnons comme domaine d'étude les recherches linguistiques exploitant un corpus textuel médiéval numérique : il peut s'agir de la BFM, ou de corpus ciblés constitués pour une étude linguistique particulière (ex.: le corpus *LEDIT*, voir (Guillot, Heiden, Lavrentiev, 2005)). Un cas un peu à part correspond aux bases

[2] Weblex est *une* implementation de la textometrie, et a ce titre fait des choix en termes d'interface utilisateur (ergonomie, scenarios), de fonctionnalites et de modelisation des corpus textuels.

associées au programme d'un concours académique (agrégation -épreuve de langue et épreuve de littérature).

Les objectifs des recherches linguistiques développées par l'équipe de la BFM depuis une dizaine d'années[3] se partagent en deux grands ensembles, quelquefois articulés : d'une part, la recherche de régularités linguistiques (captées comme ce qui s'écarte d'un comportement aléatoire, d'une combinatoire libre), en se fondant sur le volume et la systématicité des observations ; d'autre part, l'évaluation quantitative et qualitative d'un étiquetage, pour une exploitation critique et éclairée des corpus enrichis. On a donc en quelque sorte de la linguistique de corpus et de la méta-linguistique de corpus.

En ce qui concerne les recherches linguistiques, on peut encore distinguer deux types d'études.

Très classiquement, il y a la caractérisation et la modélisation de choix linguistiques : pour un objet linguistique ou une catégorie choisie, affiner sa description en précisant ses règles d'emploi contextuelles ; ou réciproquement, par l'observation et la mise en évidence des régularités contextuelles, définir des catégories plus fines. Ainsi, Guillot et Mortelmans (2008) étudient le fonctionnement du déterminant *ledit* sous ses diverses formes, dans un corpus de textes en prose des 14e et 15e siècles, en fonction d'un ensemble de paramètres morphologiques (ex. nature de la source de l'anaphore), syntaxiques (fonction, type de proposition...) et sémantiques (référent animé humain ou non), selon la théorie de l'accessibilité référentielle d'Ariel.

Un cas un peu particulier (mais central pour les corpus de textes anciens) est l'étude du changement linguistique, tel qu'il s'illustre dans le processus de grammaticalisation par exemple : en fonction de la chronologie, représentée de façon quasi continue (au fil des dates) ou cadencée (en périodes), on souhaite mettre en évidence des changements catégoriels, ou le passage d'une forme ou d'une construction dominante à une autre. Ce que l'on cherche ainsi à reconstruire, c'est la transformation du système linguistique à travers ce que l'on pourrait nommer la description du « cycle de vie » d'une forme ou d'une construction (apparition, expansion, recul ou disparition) et ses relations avec d'éventuelles variantes formelles au sein du système.

Les possibilités de synthèse statistique de l'accès Weblex à la BFM conduisent également à rechercher des corrélations entre phénomènes internes et externes. On étudie souvent le lien entre un caractère global (auteur, domaine/genre, discours direct, variété géographique...) et une fréquence de réalisation locale, selon une approche linguistique variationnelle. Par exemple, Marchello-Nizia (1997) montre que c'est dans le domaine anglo-normand que la forme *ces/cez* du démonstratif apparaît au tout début du 12e siècle, puis qu'elle s'impose rapidement au féminin et masculin pluriel comme déterminant démonstratif, ce qui neutralise deux oppositions (de genre et sémantique) qui structurent à cette même époque le système des déictiques

[3] Ces etudes mobilisent Weblex a differents degres –ce qui nous interesse ici c'est de percevoir les objectifs generaux suivis par les linguistes dans leur utilisation de la BFM en tant que corpus numerique.

du français et inaugure la spécialisation des formes caractéristiques de ce système depuis le XVIIe siècle.
Enfin, les linguistes sont intéressés à déceler des corrélations complexes, qui associent des faisceaux de traits, et montrent l'attirance ou la répulsion de tout un ensemble de traits, à la manière par exemple des études de Biber (1988). Cette démarche conduit assez naturellement à un questionnement d'ordre typologique : est-ce que mes textes, mes traits linguistiques, se regroupent en classes ? Quels descripteurs permettent de décrire les classes ? Guillot, Heiden et Lavrentiev (2007) étudient ainsi si la fréquence d'emploi du déterminant *ledit* (dans ses diverses formes) doit être corrélée à une typologie des textes selon des variables situationnelles (domaine, genre, siècle).
L'autre grand objectif actuel des études linguistiques sur la BFM est l'évaluation d'un étiquetage[4]. Il s'agit d'une part de comprendre l'assignation des étiquettes, et de percevoir la portée d'un choix de modélisation ou d'une erreur. Concrètement, l'étude d'un étiquetage prend plusieurs aspects :

- erreurs : repérer les erreurs d'étiquetage, évaluer les biais induits pour telle ou telle étude ;

- opacité : si l'étiquetage est peu ou pas documenté, comment le comprendre et restituer une définition d'usage de chaque étiquette ? comment expliquer l'origine des erreurs ?

- critique du modèle linguistique sous-jacent pour le contexte d'usage : inadéquation des catégories (par exemple la catégorie PRO:invar du TreeTagger[5], trop large), inadéquation du modèle (pour l'ancien français, pour le genre, etc.) ;

- régularité et homogénéité d'un codage manuel : même en se donnant des conventions détaillées, l'interprétation du corpus et le choix de l'étiquette peuvent varier, d'une personne à l'autre, voire pour une même personne d'un moment à un autre ;

- expressivité du formalisme d'annotation dont on dispose : par exemple il force souvent l'explicitation, pose des informations comme des évidences, introduit des contraintes pour la désignation des unités comme pour le type de relations exprimables ;

- comparaison d'étiquetages (évaluation par rapport à un étiquetage de référence, comparaison de modèles et d'outils...)

Par exemple, Guillot et Marchello-Nizia[6] rendent compte des limites actuelles de l'étiquetage morpho-syntaxique des

[4] De fait, la disponibilite et la qualite d'un etiquetage du corpus (etiquetage morphosyntaxique systematique ou marquage cible de structures pertinentes pour l'etude) conditionne ici beaucoup l'interet du travail sur corpus pour toute une gamme d'etudes linguistiques precises. Si l'etiquetage automatique est decevant, si l'etiquetage manuel est lourd, le chercheur peut hesiter a recourir a l'outillage informatique et aux textes numeriques, meme si dans l'absolu la pertinence du recours a l'outil est claire.

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/, dans une version adaptee a l'ancien francais.

[6] Recherche presentee aux Journees d'etudes du CCFM, Zurich,

démonstratifs dans les textes en ancien français par Tree Tagger : elles comparent quantitativement cet étiquetage avec celui réalisé semi automatiquement à l'aide du logiciel SATO et intégralement vérifié par des médiévistes spécialistes. Elles évaluent le taux de réussite de l'outil automatisé et font l'inventaire quantifié des erreurs, elles s'interrogent sur la manière de remédier à certaines erreurs.
Ou encore, Mortelmans et Guillot[7], mettant au point un étiquetage des déterminants de la famille de *ledit* dans la perspective de la théorie de l'accessibilité référentielle d'Ariel, sont confrontées aux questions de la régularité du codage (« faire pareil ») et aux contraintes du formalisme d'annotation (« où coder l'information : sur le déterminant *ledit* ou le nom tête, où dire qu'il n'y a pas de source à l'anaphore, que faire quand *ledit* est graphié en deux mots, comment coder le pronom zéro et où... ; comment créer le lien entre *ledit* et sa source... »).

### 2.2. Les données sur lesquelles se fonde l'enquête

Il existait déjà une archive des réflexions de l'équipe (utilisateurs de la BFM et concepteurs de Weblex) sur leur propre pratique et sur les évolutions souhaitables, sur le Wiki interne de l'équipe[8]. Ce répertoire comprend six documents au moment de la réalisation de notre enquête (nous soulignons pour chacun un mot clé qui permettra de l'évoquer commodément dans la suite de l'article) :

- des observations des pratiques et des études de besoin : une *Enquête utilisateurs* et un *Questionnaire utilisateurs* ;
- un travail sur la documentation du logiciel : un *Guide utilisateurs Weblex pour la BFM* et des propositions d'*Evolution du manuel utilisateur* ;
- des perspectives d'évolution : la proposition d'une *Maquette d'interface Weblex simplifiée pour la BFM* ; un bilan des *Etats de la BFM* (composition, évolution), avec une description des fonctionnalités attendues dans le *nouveau portail Weblex* et des différents *chantiers* programmés.

Nous avons également préparé cet article par des synthèses d'expérience présentées au séminaire d'équipe CoLiGram (octobre 2007) : *Quelques exemples d'usage de Weblex* (Céline Guillot) ; *Quelques exemples de requêtes type* (Céline Guillot) ; *Scripts et requêtes les plus fréquentes à travers Weblex* (Christiane Marcello-Nizia).
Nous disposons également d'une archive des demandes d'aide des utilisateurs et des réponses apportées, gérée par Alexei Lavrentiev, et d'un support de formation (Morinière, Heiden, 2006) inspiré des pages d'aide en ligne sur le site de la BFM.

---

octobre 2007 : « De l'usage des descripteurs : les variations d'emploi des demonstratifs suivant les caracteristiques des textes en AF (12e-13e s.) », Guillot Celine, Marchello-Nizia Christiane.

[7] Mortelmans Jesse, Guillot Celine, *Principes d'encodage pour la recherche sur LEDIT*, document de travail, mai 2006.

[8] http://weblex.ens-lsh.fr/infoling/index.php/Utilisateurs_Weblex_BFM (acces restreint).

## 3. Les pratiques observées

### *3.1. L'expression de l'objet d'étude, du motif linguistique à repérer*

Weblex permet la recherche élémentaire d'un mot par la chaîne de caractères correspondante. Mais il offre aussi une recherche par équation CQP (Christ 1994). CQP est un langage formel de requête sur corpus notamment étiquetés. Il articule trois niveaux dotés chacun d'opérateurs (joker, répétition, OU, etc.) : le niveau des occurrences (combinaison des mots), le niveau des propriétés (mobilisation possible de diverses étiquettes attachées aux occurrences), et le niveau des caractères (variations d'expression d'un mot ou plus généralement d'une valeur d'étiquette). Le langage permet également d'appliquer des contraintes liées à la structure logique (prise en compte des limites de phrases ou de paragraphes par exemple). Nous invitons le lecteur désireux d'en savoir plus à consulter une documentation du langage, par exemple dans le manuel Weblex (Heiden 2002).

Ce qui nous intéresse ici, c'est d'identifier les formes dominantes de requêtes et de trouver les raisons linguistiques qui expliquent leur expression. Nous souhaitons aussi expliciter la dynamique de constitution d'une requête complexe, en termes de stratégie de recherche. Notons que même dans un corpus étiqueté, l'expression de l'objet d'étude peut n'être pas directe, car la qualité et la pertinence de l'étiquetage sont toujours relatives. Et les stratégies d'expression d'un motif de recherche s'avèrent globalement analogues, que le corpus soit linguistiquement renseigné ou non.

#### 3.1.1. Première étape : consulter le corpus pour faire l'inventaire des formes attestées

La principale crainte du linguiste, c'est de manquer une partie significative des occurrences de son objet d'étude. En effet, les variations de tous ordres (flexionnelles, dialectales, syntaxiques, sémantiques, mais aussi orthographiques -très présentes en ancien français) démultiplient les formes de réalisation en corpus de l'objet d'étude ; et la complexité et le nombre des variantes font que leur inventaire est difficilement prévisible. Une stratégie centrale des linguistes de la BFM consiste donc à commencer par une recherche volontairement large, peu contrainte *a priori*, puis à retravailler l'inventaire des formes relevées pour cerner précisément l'objet d'étude. Ainsi, l'interface simplifiée inspirée des propositions d'Alexei Lavrentiev (*Maquette*) accompagne étroitement cette stratégie d'interrogation : recherche large, puis un affinement de la requête par retouches sur la liste des formes récoltées. Typiquement, il s'agit de pouvoir demander un paradigme via un filtre, récupérer la liste de formes correspondantes, l'ajuster.

En l'état actuel, deux fonctionnalités de Weblex répondent au besoin d'un inventaire systématique d'un ensemble de formes : l'INDEX et le VOCABULAIRE. L'INDEX permet de relever toutes les formes attestées correspondant à un certain critère : typiquement, on recherche un paradigme en utilisant les opérateurs .* pour effectuer une troncature, et l'index explicite toutes les formes correspondantes, en commençant par les plus fréquentes. Par exemple, on

demande l'INDEX de `"i?c[ie]st.*"%cd` pour obtenir des démonstratifs : *ceste*, *cest*, *cist*, *Ceste*, *cestuy*... mais arrivent aussi des intrus comme *cisterne*). Il est significatif que l'on recoure ici à une formulation du paradigme *en compréhension*, c'est à dire en donnant une règle de décision sur ce qui relève ou non du paradigme, et non *en extension*, c'est-à-dire procédant par énumération explicite des formes incluses dans le paradigme : en effet, l'enjeu est bien de limiter les *a priori* et de « laisser parler » le corpus. En ce sens aussi, l'approche est textuelle, puisqu'elle donne la première place aux attestations plutôt qu'à un lexique prédéfini externe.

| | ***Fréquence*** | ***Forme*** |
|---|---|---|
| 1 | 275 | aise |
| 2 | 31 | aaise |
| 3 | 18 | aeise |
| 4 | 11 | aisé |
| 5 | 4 | aese |
| 6 | 4 | aisse |
| 7 | 3 | aiese |
| 8 | 2 | aiesse |
| 9 | 2 | aíse |
| 10 | 1 | aaisse |
| 11 | 1 | aaissé |
| 12 | 1 | áise |
| 13 | 1 | áíse |
| | *354* | *Au Total* |

Figure1 : Exemple de présentation d'INDEX,
forme `"a[aei][aei]?.?se"%cd`

Lorsque le corpus est étiqueté, l'INDEX peut également appuyer son critère de sélection sur l'étiquetage, et demander de donner toutes les formes correspondantes ou toutes les valeurs d'une information disponible dans l'étiquetage. Par exemple, lancer INDEX de [P6="DET:demo"] permet d'obtenir les différentes formes étiquetées dans le corpus comme déterminant démonstratif : *ceste*, *ces*, *ce*, etc. Là encore, ce relevé est généralement une première étape devant être affinée.
L'usage de VOCABULAIRE, qui liste toutes les formes[9] du texte (alphabétiquement et par fréquence décroissante) est moins central, il correspond au cas où on a un objet d'étude large, sur lequel on a peu d'attentes, et un corpus pas trop gros. VOCABULAIRE est moins utilisé pour la recherche que pour l'enseignement (préparation à l'agrégation, épreuve d'ancien français), pour lequel VOCABULAIRE aide à recenser méthodiquement tous les items à étudier. Trois caractéristiques concourent à l'efficacité de cette fonctionnalité : le regroupement de toutes les occurrences d'une même forme (qui opère une *synthèse*, par rapport au texte intégral), la systématicité et la *complétude* (le relevé couvre tout le texte), et son *organisation* aux vertus heuristiques (l'ordre alphabétique rapproche des mots d'une même famille, l'ordre hiérarchique -par fréquence décroissante- rend compte des usages dominants).

[9] On peut aussi bien relever l'ensemble des mots que l'ensemble des valeurs d'une information disponible dans l'etiquetage.

3.1.2. Deuxième étape : affiner l'expression du motif recherché

Le deuxième temps de la formulation de la requête consiste à cerner plus précisément l'objet d'étude. Ce qui retient notre attention, c'est la manière dont l'usage des opérateurs des équations CQP s'adapte aux réalités linguistiques. Certains opérateurs, davantage propres aux langages formels qu'aux langues naturelles, sont totalement délaissés[10] (ou ne retrouvent une motivation que pour exploiter l'étiquetage). Quant aux opérateurs utilisés, ils le sont souvent sous des formes bien précises. Par exemple, l'opérateur d'itération (`*`) est associée au joker de caractère (sous la forme `.*`) pour couvrir des zones de variations fortes internes au mot, il n'est pas utilisé pour exprimer la répétition indéfinie d'une lettre ou d'une séquence de lettres, qui ne correspond pas au fonctionnement de la langue naturelle.

Ceci étant, la formulation la plus simple de l'objet d'étude est directement l'expression de la forme (celle de la graphie ou celle d'un étiquetage : lemme, catégorie…). Elle suppose que l'on soit dans le cas où l'unité linguistique concorde avec l'unité codée, directement accessible. On recherchera par exemple tous les éléments d'une catégorie grammaticale (notamment pour les grammèmes, d'inventaire fermé), ou bien un mot particulier par sa graphie. On peut associer une contrainte sur la graphie et une sur l'étiquette pour écarter des homographes.

Mais très souvent, la recherche se centre sur un lemme non directement identifiable par étiquetage, ou plus généralement une unité prenant plusieurs formes dans le corpus. Une grande partie des opérateurs utilisés vont être mobilisés pour capter des variations paradigmatiques. Les paradigmes peuvent être captés en compréhension (opérateurs `.*` ou `.+`) et ajustés en excluant précisément les formes indésirables repérées via l'INDEX. Par exemple:

```
[word=".+an[tsz]" &
word!="(qu|gr|dev|mainten|av|a?u?t|s|enf|neporqu)an
[tsz]"%c]
```

vise à rechercher les participes présents en sélectionnant toutes les formes qui finissent par *-ant*, *-ans* ou *-anz* et en éliminant un certain nombre de formes fréquentes qui ne sont pas des participes (*quant*, *grant*, *devant*, *maintenant*, etc.).

Les paradigmes sont aussi (et le plus souvent) captés explicitement, en extension, soit en reprenant un inventaire obtenu par une recherche en compréhension puis retravaillé (ex. pour les possessifs de troisième personne de l'ancien français : `sa|ses|sue|sues|soe|soue|soes|soues|sienne|siennes|lor|lur|leur|leurs`[11]), soit en déclinant méthodiquement plusieurs modes de réalisation du lemme (par exemple en genre et en nombre). Une requête mise au point pour une recherche sur *ledit* illustre bien ce dernier cas, car elle est structurée en quatre composantes (délimitées par des

[10] Par exemple, les operateurs s'appuyant sur l'ordre alphanumerique : le tiret pour un intervalle de caracteres (comme dans [a-d] pour noter les caracteres de a a d), et les operateurs de comparaison <, <=, >, >=. Egalement, l'operateur d'exclusion sur les caracteres (^) : dans la logique de ce que nous avons observe, l'exclusion apparait comme une facilite peu compatible avec la minimisation du silence.

[11] La forme *s'* devrait faire partie de cette liste mais est traitee de facon speciale en raison de l'apostrophe, cf. fin de la section (3.1.2).

parenthèses et séparées par des barres verticales exprimant l'alternative) qui couvrent respectivement les formes du masculin singulier, puis les formes du féminin singulier, puis celles du pluriel, enfin le cas où une segmentation différente du corpus présente la forme composée comme un seul mot[12] :

```
([word="le|du|au|ou"%c][word="-"]?[word="d[iy]c?t"]) |
([word="la"%c][word="-"]?[word="d[iy]c?te"]) |
([word="les|des|a[ul][sxz]"%c][word="-"]?
[word="d[iy]c?t?(z|s|es)"]) |
[word="(le|la|les|du|au|des|a[ul][szx]?|ou)-
?d[iy]c?t?(z|s|e|es)?"%c]
```

Comme on le voit déjà dans les exemples précédents, d'autres opérateurs sont aussi fortement mobilisés pour factoriser des variations systématiques internes au mot :

- variantes d'écriture (souvent liées à une proximité phonétique mais aussi à certaines régularités de la langue écrite ; les variantes d'écritures sont plus ou moins présentes et nombreuses selon les manuscrits et selon les éditions) (ex. `d[iy]c?te`),

- flexions « simples » (nombre, genre, cas) (ex. `afaire(s|z)`) ;

- il y a des opérateurs spéciaux pour la casse (distinction ou non des majuscules et des minuscules - `%c`) et l'accentuation (assimilation des caractères diacritiques aux caractères non diacritiques associés - `%d`) (ex. `"a[aei][aei]?.?se"%cd` pour *aise* et ses diverses formes).

Enfin, les opérateurs permettent de décrire des séquences de plusieurs mots. Cela sert lorsque l'on a affaire à une entrée lexicale composée[13] : le formalisme permet de décrire un mot suivi par un autre, ou une suite de plus de deux mots, avec de possibles insertions de mots. Les insertions sont en nombre contrôlé ou dans la limite de certaines structures (empan de trois phrases par exemple). Par exemple `"por|pour"[]{0,2}"tant"` couvre *por tant*, *por lui tant*, etc. Ces opérateurs syntagmatiques sont aussi requis pour repérer des constructions syntaxiques, par exemple, pour repérer les formes d'un objet dans un certain contexte. On a ainsi des requêtes du type :

```
[P6="DET:demo"] []* [word="mot(s)?|motz|moz|mos"]
within 2
```

(un déterminant démonstratif suivi d'une forme du lexème *mot*, éventuellement séparé du déterminant par un autre élément, par exemple *Cis joieus moz*).

Il reste le cas très particulier des caractères qui ne peuvent être utilisés tels quels pour des contraintes techniques. Du fait de son interrogation en ligne, les requêtes de l'outil Weblex sont envoyées au serveur dans le standard cgi, qui exclut des données transmises certains caractères comme l'apostrophe ou le point-virgule. Les requêtes qui doivent porter sur l'un de ces caractères doivent alors soit désigner ces caractères par leur code octal (tel que `\\\047` pour

[12] Le resultat de cette requete, couvrant 111 variantes graphiques dans le corpus etudie, est publie dans (Guillot, Heiden, Lavrentiev 2007).

[13] Le caractere compose ou non d'une unite est relatif au codage du corpus vu par l'outil textometrique : en effet, on peut avoir opere un codage qui represente les locutions et d'autres formes composees comme des unites.

l'aostrophe), soit recourir à des heuristiques de contournement de la difficulté, par exemple :

```
[word="l[^a-z]"%cd]
```

pour obtenir l'article élidé *l'* ou *L'*, ici exprimé comme la lettre *L* (majuscule ou minuscule) suivi d'un caractère qui n'est pas une lettre. Ces biais technologiques sont très pénalisants pour les utilisateurs non avertis, qui ne comprennent pas d'où vient l'échec de leur requête.

La possibilité de repérage de motifs linguistiques précis ouvre alors deux grandes perspectives d'investigation, que nous étudions dans les deux sections suivantes.

### *3.2. Observation en contexte des structures linguistiques*

Le corpus permet de rassembler toutes les formes attestées et tous les contextes d'usage d'un motif donné, et de déceler des régularités de construction et des affinités. C'est ainsi que l'on se donne des moyens pour qualifier syntagmatiquement un objet d'étude, et souvent pour distinguer et catégoriser plusieurs types sous-jacents à une même forme de surface.

Là encore, les choix de conception de l'interface simplifiée BFM sont révélateurs : le premier champ (en haut à gauche) est celui d'entrée du motif, introduit par le terme « Rechercher : » ; et le bouton pour lancer le traitement s'appelle « Lancer la recherche ». Autrement dit, la consultation et les traitements textométriques sur le corpus sont essentiellement pensés comme des recherches d'occurrences (plutôt que des calculs de synthèse statistiques), occurrences à visualiser en listes (notamment au moment de la mise au point de l'expression de la recherche) puis dans leur contexte (Morinière, Heiden, 2006).

Weblex offre deux modes d'affichage des passages où se réalise le motif recherché : la fonction CONTEXTE et la fonction CONCORDANCE[14]. Ces deux fonctions partagent les mêmes possibilités de paramétrage (réglage de la taille du contexte de part et d'autre de l'occurrence, tri des contextes suivant plusieurs critères) et apparaissent simplement comme deux formes de présentation différentes.

passion, v.47 : gran e petit Deu van laudant . Ensobre tot petiz **enfan** osanna semper van clamant . A la ciptad cum aproismet , et el la vid e ·lla' sgarded , de son piu c

passion, v.61 orn t' arberjaran et a terra crebantaran . Los tos **enfanz** , a males penas aucidront ; en tos belz murs , en tas maisons pedra ·ssubr' altre non laiseront . L

passion, v.378 mne non a vertud . Etqui era li om primers e·l soi **enffant** per son pechet ; e li petitet e li gran etqui estevent per mulz anz . Quar anc non fo nulom carnals

Figure 2 : Exemple de présentation de CONTEXTE

[14] La fonction INDEX peut egalement servir a relever les differents contextes immediats d'un motif suivant une structure determinee, par exemple le motif avec le mot qui precede et les deux mots qui le suivent. Cet usage particulier est peu connu.

| | | | | |
|---|---|---|---|---|
| 1 | passion , v.47 | : gran e petit Deu van laudant . Ensobre tot petiz | **enfan** | osanna semper van clamant . A la ciptad cum aproismet , et el la vid e ·lla' sgarded , de son piu c |
| 2 | passion , v.61 | orn t' arberjaran et a terra crebantaran . Los tos | **enfanz** | , a males penas aucidront ; en tos belz murs , en tas maisons pedra ·ssubr' altre non laiseront . L |
| 3 | passion , v.378 | mne non a vertud . Etqui era li om primers e·l soi | **enffant** | per son pechet ; e li petitet e li gran etqui estevent per mulz anz . Quar anc non fo nulom carnals |

Figure 3 : Exemple de présentation de CONCORDANCE

CONTEXTE fournit une liste des contextes, sous la forme d'une suite d'alinéas, introduits chacun par une référence indiquant la localisation de l'extrait dans le corpus. L'occurrence du motif recherché est affichée en gras. On a donc une présentation visuellement proche de celle d'un texte. En revanche, CONCORDANCE se présente plutôt comme un tableau dans lequel le texte est réparti dans trois colonnes : au centre, la réalisation du motif recherché, et de part et d'autre ses contextes gauche et droit.

Ces différences de présentation sont-elles sensibles pour les linguistes ? Et si oui, qu'est-ce qui motive le choix de l'une ou de l'autre ? Notre enquête révèle que les deux fonctionnalités sont utilisées, et le sont de façon effectivement différenciée. C'est essentiellement le mode d'exploitation du résultat qui oriente le choix : la présentation CONTEXTE, plus classique et plus synthétique, convient davantage à une impression, pour poursuivre l'analyse sur le papier, et notamment pour communiquer cette liste d'extraits dans un support de cours ou dans une publication. La présentation CONCORDANCE, plus aérée et plus structurée, est plus propice à une consultation à l'écran, avec sa dynamique (multiples possibilités de tris, liens hypertextes permettant d'accéder à un contexte plus large, etc.). Elle se prête également à un export dans un tableur, offrant encore d'autres possibilités de mise en forme et de consultation dynamique (filtres, etc.).

Les deux fonctionnalités sont finalement en pratique distinguées par des types d'usage (lecture sur papier et édition de documents *vs* consultation numérique avec dynamiques de réorganisation) plutôt que par des propriétés de fond[15]. Un linguiste peu habitué à travailler à l'écran prend ainsi l'habitude de recourir exclusivement à la fonction CONTEXTE. Un linguiste familiarisé avec les deux types de supports jongle avec les deux fonctions.

Lorsqu'on les consulte, les linguistes soulignent spontanément l'importance pour eux du réglage souple de

[15] Et en effet, le mode de realisation des concordances mis en avant dans Weblex (parametrage par defaut) et dans l'interface simplifiee est celui dit « en tableau » : des que le contexte depasse quelques mots, il est presente sur plusieurs lignes, ce qui altere les effets de superposition obtenus par les tris. Or ces effets sont un des atouts fondamentaux des concordances, qui permettrait une distinction de fond entre les concordances et d'autres formes de presentation de contextes (Pincemin et al., 2006).

la taille des contextes. Il s'agit d'abord évidemment de traiter séparément le contexte gauche et le contexte droit, qui jouent généralement des rôles très différents du fait de la structuration syntaxique. Et il s'agit également de pouvoir accéder aussi bien à des contextes courts (focalisés sur les occurrences dans un même composant syntaxique, par exemple) qu'à des contextes élargis (lorsque la portée de l'observation doit dépasser la phrase, par exemple dans le cas de relations anaphoriques). Si bien souvent « trois phrases maximum suffiraient », on préfère néanmoins avoir accès à une expansion du contexte « à volonté »[16] (*Enquête*). L'argument linguistique est que la taille pertinente pour le contexte ne peut être fixée dans l'absolu et dépend du type de phénomène observé et du genre des textes. Il y a aussi un argument plus technique, lorsque l'extraction de contextes est une étape dans une chaîne de traitements et sert à constituer un sous-corpus : là encore, c'est la perspective de constitution et d'utilisation du sous-corpus qui fixe la taille des contextes extraits, potentiellement larges. Ainsi, la fonctionnalité textométrique de retour au texte[17], qui pourrait prendre avantageusement le relais d'une fonction CONTEXTE lorsque la taille des contextes s'élargit et provoque de multiples recouvrements, ne démotive pas complètement des relevés de contextes très étendus.

La taille du contexte serait exprimable de différentes façons. Le paramétrage de CONCORDANCE et CONTEXTE donne un réglage en nombre de caractères. Mais CQP permet également l'extraction de passages autour d'un motif, qui peuvent s'exprimer en nombre de mots et en nombre de phrases[18]. Ces dernières formes de contextes sont naturellement plus linguistiques. Ceci étant, la délimitation en nombre de caractères reste heuristiquement efficace, et n'est pas dépourvue de sens en ce qu'elle reflète aussi une certaine distance syntagmatique.

### *3.3. Décomptes contrastifs et répartition*

L'identification de toutes les occurrences d'un item linguistique en corpus permet d'en relever les différentes formes attestées et d'observer leur répartition sur différentes parties du corpus. Dans le cadre de la BFM, on cherche par exemple à en tirer des enseignements sur les possibles linguistiques et sur les usages à une période donnée ou

[16] Dans Weblex les contextes peuvent etre etendus sans limite *a priori*, ils sont dans les faits bornes par les capacites du navigateur et du serveur, ce qui n'a pour le moment jamais gene les etudes linguistiques. Depuis 2008, pour la BFM, une limite plus restrictive de 1500 caracteres de part et d'autre du motif recherche est desormais imposee par les editeurs des editions de reference utilisees.

[17] La consultation et le feuilletage du texte avec mise en valeur des occurrences d'un motif est implementee dans Weblex mais n'est en fait pas disponible dans la BFM, pour des raisons de droits de propriete (limitation de la taille des citations d'un texte). Pour d'autres corpus medievaux consultables par Weblex (par exemple la *Quete du Graal*), on a l'acces a l'edition et a l'image de la page du manuscrit.

[18] La mesure en phrases est accessible (operateur `expand to`) dans la mesure ou les phrases ont ete prealablement codees dans le corpus. Quant aux mots, ils sont definis ici par la segmentation en occurrences egalement fixee au moment de l'integration du corpus dans l'outil ; on compte alors un nombre d'occurrences en utilisant le joker d'occurrence `[]`, soit repete, soit muni d'indications sur le nombre de repetitions, par exemple `[]{1,3}` signifiant « 1 a 3 mots quelconques ».

pour un type de texte donné. Le linguiste prête alors attention aux fréquences, lorsqu'elles révèlent des dominances et des affinités. Il peut s'appuyer sur elles pour reconstruire des formes à partir des multiples occurrences et subdivisions du corpus, comme l'allure de l'évolution d'un motif ou la délimitation d'un domaine d'usage (Marchello-Nizia 2004).

La fonction INDEX opère un relevé de toutes les formes de réalisation attestées pour le motif recherché, en indiquant la fréquence de chaque forme (des fréquences les plus élevées aux fréquences les plus faibles) et la fréquence totale du motif. Si le corpus est partitionné (c'est-à-dire si l'on prend en compte un fractionnement du corpus en plusieurs parties, par exemple en époques ou en genres textuels) la fonction INDEX opère ce relevé pour chacune des parties, mettant ainsi en avant les formes qui dominent dans chaque partie, et -ce qui est surtout utilisé- la variation de la présence du motif (toutes formes confondues) au fil des parties.

Pour pouvoir comparer les fréquences d'une partie à l'autre, les linguistes rapportent la fréquence à la taille des parties, obtenue par la commande DIMENSIONS. Ainsi, l'analyse se base sur le pourcentage d'occupation de la partie par le motif, ou encore sa fréquence relative (fréquence du motif pour 10 000 occurrences). Par exemple, la présence des formes du déterminant *ledit* croît du 13e au 15e siècle : réservées au départ au domaine juridique (fréquence relative de 23, alors que la fréquence relative est proche de zéro dans les autres domaines), en deux siècles elles gagnent sensiblement les autres domaines (avec des fréquences relatives de 90 pour le domaine historique, de 23 pour la didactique, de 6 pour la littérature) (Guillot, Heiden, Lavrentiev 2007).

Dans la même perspective d'études contrastives, la fonction REFERENCES génère un index (au sens usuel du mot) pour les différentes attestations du motif : elle produit un relevé alphabétique des formes attestées, avec pour chacune la liste de ses localisations, sous la forme par exemple du nom du texte et de la page où figure l'occurrence trouvée. Elle est quelquefois mobilisée en complément à INDEX, pour affiner la vision de la répartition, « quand les textes sont nombreux à donner une forme ».

## 4. Ecarts entre les pratiques des linguistes et les possibilités d'analyse textométriques

### *4.1. Typologie des écarts*

La synthèse que nous venons de faire des pratiques des linguistes utilisateurs de la BFM montre, pour un connaisseur des outils textométriques et en particulier de Weblex, que les possibilités de l'outil sont loin d'être toutes exploitées.

Nous relevons tout d'abord des fonctionnalités dont les linguistes expriment le besoin et regrettent de ne pas disposer, alors qu'elles existent dans l'outil. Cela concerne le paramétrage (non élagage des hapax et des formes outils ; affichage des résultats dans une autre fenêtre) mais aussi les traitements disponibles (les informations sur le corpus sont consultables par la fonction BIBLIO, la

restitution du texte peut être soignée dans la fonction EDITION). Cet écart entre les possibilités et les usages s'explique d'abord tout bonnement par la non disponibilité de certaines fonctionnalités de Weblex pour le corpus BFM, du fait de sa composition et de sa réalisation (problèmes de droits de propriété notamment). Un autre facteur d'écart est une connaissance incomplète de l'outil : on l'utilise pour quelques fonctionnalités connues et satisfaisantes, sans forcément en avoir fait le tour. De fait, les possibilités de réglages et de commandes sont très nombreuses, et celles qui auraient pu retenir l'attention du linguiste sont quelquefois tout simplement « noyées » au milieu d'autres. L'interface simplifiée est une manière de répondre à cette difficulté.

Le cas le plus remarquable de ce type d'écart (fonctionnalités existantes mais méconnues) est celui de la fonction SPECIFICITES. En effet, cette fonctionnalité réalise un calcul qui a été conçu précisément pour répondre à la préoccupation du linguiste, lorsqu'il est désireux de repérer des contrastes d'usage significatifs entre des parties de tailles différentes (Lafon 1980). La fonction SPECIFICITES correspondrait donc *a priori* pleinement aux études du type de celles décrites en partie 3. Comment expliquer alors sa désaffection ? Pas par simple ignorance ni par désintérêt de principe ; ce serait plutôt lié la question simple mais fondamentale d'une bonne compréhension et pleine appropriation du calcul. Il faut comprendre le calcul pour être capable de l'utiliser de façon maîtrisée et juste, pour savoir interpréter le résultat et lui trouver une signification. Autrement dit, nous percevons là une exigence éthique du chercheur, qui se refuse à adopter un traitement certes sophistiqué mais dont la signification lui échappe. Pour le linguiste non formé aux techniques de la statistique textuelle, le calcul de pourcentages ou de fréquences relatives[19] est parlant ; alors que la significativité d'un indice de spécificité lui est opaque.

Il y a aussi le cas où les linguistes sont intéressés par certaines techniques statistiques qui ne sont pas disponibles dans Weblex (par exemple des analyses factorielles, telles que pratiquées par Biber, ou encore des classifications), mais pour lesquelles ils n'évoquent pas leur besoin d'en disposer dans Weblex, soit qu'ils ne perçoivent pas que ces techniques relèvent du même domaine, soit qu'ils pressentent la technicité de ces calculs et ne se sentent pas prêts à les mettre en œuvre.

Enfin, tout calcul proposé par Weblex n'est pas nécessairement pertinent pour les études réalisées dans le cadre de la BFM : « Weblex est un outil magnifique [...]. Mais certaines fonctionnalités ayant été introduites pour de la lexicométrie d'un certain type, il ne faut pas se tromper et les utiliser sur notre corpus ». C'est aussi vrai pour les paramétrages par défaut, et notamment pour les seuils, qui sont à redéfinir selon l'approche linguistique adoptée. Ici, nous avons vu que les études sur la BFM adoptaient généralement une approche lexicale morphosyntaxique, focalisée sur un objet d'étude dont on caractérise les variations en contexte et selon différents paramètres

[19] Ces calculs sont accessibles via un export de Weblex vers d'autres logiciels.

externes (époque, genre textuel, etc.). Cela explique le moindre intérêt de :

- caractérisations globales (non focalisées sur un motif) comme celles données par les fonctions ZIPF[20], PARETO[21], LONGUEUR DES PHRASES[22] ;

- démarche exploratoire, sans *a priori* sur les formes à étudier (*vs* observation d'un item précis) : REPARTITIONS[23], SEGMENTS REPETES[24], COOCCURRENTS[25] ;

- autres applications issues d'autres domaines : par exemple TERMES correspond à une application d'extraction de candidats termes pour des terminologues.

Au compte des fonctionnalités de Weblex non exploitées par les linguistes de la BFM, on trouve encore :

- REPARTITION[26], qui suppose de fait un corpus à linéarité interne, type intratextuel (cas typiquement d'un corpus composé d'un seul texte long) ; de plus le caractère cumulatif de la courbe produite ne correspond pas aux attentes des utilisateurs de la BFM ;

- LEXICOGRAMME[27], sans doute plus thématique que morphosyntaxique, quoique les relations mises en évidence soient davantage morphosyntaxiques si on limite le calcul à un contexte étroit (dm petit) ; mais la concordance reste sans doute plus appropriée ;

- LEXICOGRAMME(S) RECURSIF(S)[28] : plus global (étude élargie plutôt que focalisée).

---

[20] La fonctionnalite ZIPF de Weblex donne la gamme des frequences realisees dans le corpus, de la plus elevee (celle du mot le plus frequent) a la plus faible (1, pour les hapax, a savoir les mots qui n'apparaissent d'une seule fois) ; chaque frequence est accompagnee du nombre de mots comptes avec cette frequence.

[21] La fonctionnalite PARETO de Weblex trace le diagramme caracteristique de la gamme des frequences, mettant en evidence la loi dite de Zipf-Pareto, a savoir que le produit de la frequence d'un mot par son rang (selon un classement par frequence decroissante) est constant.

[22] Histogramme des longueurs des phrases du corpus.

[23] La fonctionnalite REPARTITIONS de Weblex s'appuie sur une mesure statistique de la regularite d'apparition d'un mot au fil du texte ; elle liste les mots du texte en commencant par ceux ont la repartition est la plus irreguliere (dite « en rafales »).

[24] La fonctionnalite SEGMENTS REPETES recense les suites de mots recurrentes, en commencant par les plus remarquables (les plus longues et les plus frequentes) ; elle permet notamment de reperer des figements ou des delimitations de mots inapropriees.

[25] La fonctionnalites COOCCURRENTS de Weblex recense les couples de mots qui s'attirent mutuellement dans un certain voisinage (par defaut une distance moyenne de mille mots) ; elle les liste en commencant par les attirances les plus fortes.

[26] La fonctionnalite REPARTITION de Weblex trace la courbe de la frequence cumulee d'un motif au fil du corpus.

[27] Le LEXICOGRAMME est un affichage sous forme de deux listes des principaux cooccurrents avant et apres le mot considere.

[28] Les LEXICOGRAMMES RECURSIFS affichent des graphes des principales cooccurrences dans un texte, a partir d'un mot ou dans leur globalite.

### *4.2. Interprétation des écarts*

Tout d'abord, comme nous avons eu l'occasion de le souligner à propos du non usage de la fonction SPECIFICITES, il faut comprendre pour utiliser. Si la signification profonde d'un calcul lui échappe, le chercheur préfère recourir à des heuristiques plus claires pour lui, et qu'il saura mieux interpréter.
Egalement -et cela a déjà été perçu et explicité par l'équipe- les utilisateurs de la BFM disposent du manuel de référence de Weblex (Heiden 2002), mais il manque un guide utilisateur. Un tel guide compléterait le manuel de référence en développant davantage des composantes pragmatiques, didactiques et méthodologiques :

- pragmatique :
  - l'exposé est construit à base d'exemples ;
  - il commence par des usages dominants (plutôt que de parcourir les fonctionnalités selon une logique théorique) ;
  - il décrit des cas particuliers utiles (par exemple la manière de rechercher un motif comportant une apostrophe) ;
- didactique
  - l'initiation au fonctionnement de l'outil gagne à s'appuyer sur la connaissance d'outils proches et connus du lectorat visé, tels que Frantext ;
  - un effort particulier doit être fait pour adopter une terminologie très accessible ;
- méthodologie
  - le guide présente non seulement les fonctionnalités, mais donne des stratégies de mise en œuvre : pour la BFM, on a par exemple la stratégie de mise au point de l'expression du motif recherché, selon laquelle on commence par une requête trop large que l'on affine ensuite ;
  - le guide devrait également apporter des éléments de réponse pour des difficultés typiques, par exemple le cas de résultats volumineux.

Bien entendu, des actions de formation (séminaires, école d'été, etc.) iraient dans le même sens.
Enfin, une fonctionnalité encore méconnue peine à être adoptée car le pionnier qui l'utilise s'isole au plan des pratiques de sa communauté de recherche : ses résultats sont difficilement comparables, et moins réexploitables par d'autres (Marchello-Nizia 2004).

### *4.3. Enseignements et perspectives*

En pratique, les traitements textométriques doivent s'intégrer dans une pratique qui inclut d'autres types d'analyse et qui surtout doit pouvoir être diffusée par des publications et des enseignements. Ainsi, les possibilités d'export ne sont pas une option, mais peuvent être décisives dans l'adoption de l'outil. Nous avons noté par exemple que ce sont les types d'export possibles qui souvent arbitraient les usages de CONTEXTE ou de CONCORDANCE, en tirant profit de leur compatibilité avec les outils de bureautique usuels (traitement de texte, tableur). Dans le même ordre d'idée, si le traitement peut

s'accompagner de modes de visualisation (histogrammes, courbes, etc.), il est important que ces représentations puissent être exportées pour réutilisation dans d'autres supports.
L'interface de l'outil textométrique -ici Weblex- gagne à se spécialiser : c'est ce qui est expérimenté maintenant depuis 2006. Moins riche (mais l'accès à l'interface complète est toujours possible), l'interface dite simplifiée est surtout moins lourde et plus claire. Par exemple, au lieu d'avoir pour le motif à rechercher trois zones de saisie introduites par « source A », « source B », « source C » (ces deux dernières n'étant en fait requises que pour certains calculs particuliers donnant la possibilité d'afficher simultanément plusieurs résultats), on a une seule zone de saisie introduite par « Rechercher ». Le paramétrage par défaut est à adapter à la communauté des utilisateurs (par exemple pour la BFM, par défaut, il ne faut pas distinguer la casse, ni élaguer les basses fréquences ou les mots grammaticaux) ; mieux, il faudrait que chaque utilisateur puisse se définir un profil de paramétrage par défaut enregistrable (voire plusieurs).
La syntaxe du langage CQP est très riche, et donc difficile à mémoriser. Plusieurs aides pourraient faciliter son utilisation :

- la mise à disposition de fiches aide-mémoire en ligne : récapitulatif des éléments de la syntaxe ; exemples correspondant à des usages fréquents ;

- la sauvegarde de requête, accompagnée d'une glose (nom explicatif, commentaire). La requête est alors réutilisable comme sous-requête, pour la mise au point de requêtes élaborées ; elle pourrait éventuellement être partagée dans une communauté d'utilisateurs[29]. Elle peut servir d'exemple adaptable, de « moule » (il est plus simple de modifier une requête existante que d'en concevoir une entièrement nouvelle) ;

- le langage CQP pourrait en partie être interfacé de façon graphique. Déjà, dans la maquette d'interface simplifiée pour la BFM, la prise en compte ou non de la casse ou des diacritiques serait activable par des boutons ;

- enfin, il faudrait réfléchir à l'intérêt de concevoir un langage d'interrogation plus adapté à la langue naturelle que les expressions régulières, en prenant en compte les observations faites en partie 3.1 : l'étoile (*) est essentiellement utilisée pour représenter une forme de troncature, ou une répétition qui pourrait (ou devrait) en fait être bornée ; les alternatives de caractères correspondent très souvent à des variantes phonétiques (dont certaines fréquentes, comme *i/y*, *c/ss*, *ai/e*), ou à des catégories de type alphanumérique, ponctuation, séparateur, etc. Il s'agirait peut-être de proposer de nouveaux opérateurs, voire de donner à l'utilisateur la possibilité d'en définir.

[29] Cela correspondrait a une pratique de fait, l'echange de fichiers de requetes.

Un dernier grand trait des besoins des linguistes concerne la dynamique de la recherche : si l'on définit une partition du corpus (c'est-à-dire un découpage en parties, par exemple en genres ou en périodes), l'analyse conduit souvent à remodeler le découpage (typiquement sortir ou ajouter un texte). Quelquefois, on voudrait pouvoir focaliser l'étude sur une partie de corpus, voire sur des extraits de textes. Il y a aussi une dynamique d'élaboration des requêtes : il s'agirait d'accompagner l'élaboration des requêtes complexes, et l'enregistrement de requêtes sur lesquelles prendre appui pour aller plus loin. Et il y a une dynamique de l'annotation et de l'édition du corpus : l'analyse conduit à corriger et à enrichir les unités linguistiques et textuelles. Les perspectives sont vertigineuses, et appellent la mise au point de méthodologies nouvelles.

## 5. Conclusion

L'objet de cet article pouvait surprendre *a priori* : y a-t-il quelque chose à tirer d'une observation des usages d'un outil d'interrogation de corpus par une communauté de linguistes ? y a-t-il quelque fait ou comportement qui ne soit évident, connu ? cela peut-il nous apprendre du nouveau au plan de la linguistique, y a-t-il matière pour une publication scientifique en linguistique ? A l'issue de cet article, nous pouvons faire le point des résultats qui ont pu être tirés de notre enquête.
Par delà la contingence d'un outil particulier et d'une base de textes donnée dans un certain état d'avancement, plusieurs attentes fortes des linguistes en matière de travail sur corpus se dégagent :

- pouvoir observer en corpus toutes les attestations d'une unité linguistique, *y compris dans ses formes peu fréquentes ou déviantes* ; cf. non élagage par défaut[30], stratégie de mise au point d'un motif de recherche paradigmatique en compréhension puis paradigmatique en extension ;

- importance de la *contextualisation*, tant intratextuelle à plus ou moins longue portée qu'intertextuelle, et recherche de régularités dans les *affinités* entre la forme de réalisation d'un item linguistique et un certain type de contexte.

Les recherches linguistiques autour de la BFM sont pratiquement toujours focalisées (sur un élément morphologique, lexical, typographique, syntaxique,...) ; on ne s'étonnera donc pas de voir assez systématiquement délaissées les procédures textométriques mises au point pour des explorations de corpus globales, non focalisées, et qui sont davantage mobilisées et appréciées dans d'autres domaines (analyse du discours par exemple). Se dessine donc une opposition généralement peu perçue, mais bien réelle, entre des procédures d'analyse focalisées et des procédures panoramiques (non orientées par un item linguistique particulier).
Nous avons été amenés à souligner l'importance de reconnaître l'ancrage de l'outil dans une démarche globale.

[30] On travaille sur *tous* les mots du corpus, on considère qu'il n'y a pas de « mots vides » ou « mots outils » à écarter de l'analyse.

Celle-ci intègre le choix critique d'une édition du texte et l'établissement d'un corpus, la multiplicité des paramètres à considérer, la révision méthodique des hypothèses et l'évolution de l'analyse, l'intérêt heuristique, rhétorique et pédagogique des visualisations graphiques, la reprise des résultats (publication, support de cours, autre logiciel, etc.).
L'interprétabilité des calculs est essentielle : les techniques les plus simples (à base de tris, de rapprochements visuels, d'analogie avec les modes de représentation et d'expression courants), et qui ont été retenues par la textométrie parce qu'elles correspondent à des heuristiques efficaces, ont de ce fait un avantage décisif. La formation et la réalisation d'un guide de l'utilisateur ouvrent des perspectives pour la familiarisation des linguistes avec des techniques plus statistiques, mais la présentation des techniques de base et de leur mise en œuvre reste capitale.
Nous avons aussi observé que le langage formel des expressions régulières est expressif, mais pas particulièrement adapté à la morphologie des langues naturelles : peut-être que d'autres formes d'expression d'un motif à rechercher seraient à proposer complémentairement, qui soient plus proches des régularités linguistiques (articulation en morphèmes et système des flexions et dérivations, paradigmes de variation stabilisés pour un contexte donné, répétitions faibles, etc.). D'une manière générale, un équilibre est à trouver entre la puissance des traitements dans toute leur généralité, et leur sélection et mise en relief dans le contexte d'une communauté de pratique, et d'usages personnels. Ainsi en va-t-il du langage d'interrogation, de la présentation de l'interface, des paramétrages par défaut. Cet équilibre entre généralité et spécialisation ne serait pas à chercher en termes de compromis, mais d'articulation entre des traitements généraux, complets et puissants, toujours disponibles, et des profilages contextuels à forte pertinence.

**Références bibliographiques**


Biber Douglas (1988), *Variation across speech and writing*, Cambridge University Press.

Christ Oliver (1994), « A Modular and Flexible Architecture for an Integrated Corpus Query System ». in *Proceedings of COMPLEX'94*, 3rd Conference on Computational Lexicography and Text Research, Budapest, Hungary, July 7-10, 1994, p. 23-32.

Dendien Jacques (2002),« Une théorie des objets textuels et des moteurs de recherche dans les bases textuelles », in *Actas del secundo seminario de la escuela interlatina des altos estudios en lingüistica aplicada Matematicas y tratamiento de corpus*, Logroño, Fundación San Millán de la Cogolla, p. 85-93.

Guillot Céline, Heiden Serge, Lavrentiev Alexei (2007), « Typologie des textes et des phénomènes linguistiques pour l'analyse du changement linguistique avec la Base de Français Médiéval », in *Linx*, numéro hors série *Corpora et Questionnements du littéraire*, Denise Malrieu (dir.), p. 125-139.

Guillot Céline, Mortelmans Jesse (2008). « Clarté ou vérité, *ledit* dans la prose de la fin du Moyen-âge », in

*Discours, diachronie, stylistique du français. Etudes en hommage à Bernard Combettes*, O. Bertrand et al. (dir.), Bern, Peter Lang, p. 307-323.

Heiden Serge (2002), *Weblex. Manuel Utilisateur. Version 4.1 (janvier 2002)*, Lyon, Laboratoire ICAR UMR 5191 CNRS & Université de Lyon, 180 pages. En ligne : http://weblex.ens-lsh.fr/doc/weblex.pdf

Heiden Serge, Lavrentiev Alexei (2004), « Ressources électroniques pour l'étude des textes médiévaux : approches et outils », *Revue Française de Linguistique Appliquée*, IX, 1, *Linguistique et informatique : nouveaux défis*, Benoît Habert (dir.), Amsterdam, De Werelt Eds, p. 99-118.

Lafon Pierre (1980), « Sur la variabilité de la fréquence des formes dans un corpus », *M.O.T.S*, 1, p. 127-165.

Lecomte Josette (2002), *La Base FRANTEXT. Documentation de l'utilisateur*, Nancy, ATILF, 76 pages. En ligne : http://www.atilf.fr/atilf/produits/Tuturiel_Frantext.pdf

Marchello-Nizia Christiane (1997), « Variation et changement, quelles corrélations ? », *Langue française*, 115, p.111-124.

Marchello-Nizia Christiane (2004), « Linguistique historique, linguistique outillée : les fruits d'une tradition », *Le français moderne*, 72, 1, *Traitement automatique et ressources numérisées pour le français*, Catherine Fuchs et Benoît Habert (dir.), p. 58-70.

Morinière Mélanie, Heiden Serge (2006), Manuel de formation à l'utilisation de la Base de Français Médiéval (BFM) et à la syntaxe du moteur de recherche Weblex, Equipe du projet BFM, ENS-LSH, Lyon, octobre 2006, 6 pages.

Pincemin Bénédicte (2006), « Concordances et concordanciers -De l'art du bon KWAC », in *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation*, Actes du 17e Colloque d'Albi Langages et Signification (Albi, 10-14 juillet 2006), Carine Duteil-Mougel et Baptiste Foulquié (dir.), p. 33-42 ; et *Texto!*, XI, 2 [en ligne http://www.revue-texto.net/ ISSN 1773-0120].