Pour une exploration humaniste des textes : AnaLog

Marie-Hélène Lay [1], Bénédicte Pincemin [2]

[1] Université de Poitiers – Laboratoire FoReLL – MSHS de Poitiers –
95 av. Du Recteur Pineau 86000 Poitiers – France

[2] CNRS – Université de Lyon – Laboratoire ICAR– ENS-LSH – 15 parvis René Descartes–
BP 7000 – 69642 Lyon cedex 7 – France

Résumé

Dans ses fonctionnalités d'exploration de corpus annotés, AnaLog est un outil de textométrie. Il permet la lecture linéaire ou synthétique de textes annotés, les inventaires et décomptes d'unités linguistiques, leur contextualisation simultanée sur les axes paradigmatiques et syntagmatique. Pour répondre aux besoins d'utilisateurs humanistes et de linguistes, AnaLog a été conçu spécialement pour interroger simultanément des observables construits (corpus de textes) et des modèles de représentation (ressources d'annotation), afin d'étudier le résultat -toujours recomposable- issu du croisement des deux : les corpus annotés. L'originalité d'AnaLog tient aussi aux modes d'interrogation et de parcours des données et des résultats, ancrés sur la visualisation constante, sous forme de tableaux, des données textuelles et des paramètres actifs.

Abstract

In regard to its annotated corpus exploration capabilities, AnaLog is a textometric tool. It allows linear and synthetic reading of annotated texts, inventory and counting of linguistic units, and their simultaneous contextualization on the paradigmatic and syntagmatic dimensions. To meet the needs of humanist users and linguists, AnaLog was designed to explore, at the same time, selected raw contents (text corpora), and their corresponding model-dependent representations (annotation resources), and then to dynamically analyze the composition of these two dimensions: annotated corpora. AnaLog is also singular by its unique navigation mode within data, results and parameters, using simple tables that can be sorted, filtered and modified to refine the current analysis.

Keywords: software for textual data analysis, textometry, digital humanities, corpus annotation, user interface, linguistic analysis, syntagmatic and paradigmatic views, concordance, corpus and quantitative linguistics, linguistic pattern searching, intuitive querying

1. Contexte d'usage et principes de conception

1.1. Outiller une lecture savante dans le domaine des humanités

Le logiciel AnaLog est développé dans le cadre des BVH, les Bibliothèques Virtuelles Humanistes [1]. Cette bibliothèque numérique intègre divers outils permettant l'exploitation des ouvrages qu'elle diffuse : les BVH veulent proposer à leurs visiteurs des modes d'accès ne se limitant pas aux données bibliographiques et à la recherche sur le texte intégral.

[1] http://www.bvh.univ-tours.fr.

Les outils mobilisés sont issus en particulier des communautés du traitement d'images (OCR, découpage de pages, identification de motifs graphiques) (Ramel et al., 2006), du traitement automatique des langues (TAL) et de la textométrie (moteurs de recherche, ressources linguistiques, annotation, concordanciers). Ainsi, les BVH se conçoivent comme un espace de *lecture outillée* (Demonet and Lay, 2008), permettant d'utiliser des corpus annotés et des ressources pour leur description et leur analyse (telles que des dictionnaires). Le corpus dont il est question ici, essentiellement des textes du 16° siècle se trouve depuis longtemps déjà associé aux recherches sur ce type d'outils : les premières mises à disposition de l'œuvre de Rabelais sous Hyperbase (Brunet, 2001) remontent à 1994, 1995 pour la version en ligne (Demonet, 1998). Puis, en 1999 avait été développé un taggeur lemmatiseur pour les textes hétérographiés, Humanistica (Antoni-Lay and Demonet, 2000). AnaLog vient ensuite [2]. Il s'agit en particulier d'utiliser des briques logicielles pour analyser des corpus à des fins d'études et d'analyses linguistiques et littéraires : caractérisation d'auteurs, d'œuvres, de discours, de langues ou d'états de langue.

Dans ses fonctionnalités d'exploration de corpus annotés, AnaLog est un outil de textométrie. Il permet la lecture linéaire et synthétique de textes annotés, les inventaires et décomptes d'unités linguistiques, leurs contextualisations. Son originalité tient essentiellement à deux aspects, présentés dans les deux sections suivantes. Le premier est l'ergonomie centrée sur les textes, notamment pour lancer les traitements. Le second aspect original, est le travail non seulement sur le corpus, mais aussi sur les ressources descriptives ; car AnaLog permet d'interroger simultanément, en les rendant visuellement accessibles, des observables construits (corpus de textes – ordre syntagmatique) et des modèles de représentation (ressources d'annotation – ordre paradigmatique), afin d'étudier le résultat – toujours recomposable – issu du croisement des deux : les corpus annotés.

1.2. Travailler constamment à partir des textes : une ergonomie appropriée au mode de travail des Humanistes

Le contexte de documentation et d'étude savante dans le domaine des humanités met clairement, d'entrée de jeu, le texte et le lecteur au centre de la problématique. Les choix d'implémentation partent de cette posture : (1) il faut partir des textes et autres ressources disponibles (informations linguistiques ou autres types de méta-données), plutôt que de la formulation de requêtes et du lancement de commandes à appliquer au texte ; (2) il faut aussi prévoir les opérations concrètes que l'on souhaite pouvoir réaliser depuis chaque type d'état des données (le texte initial, une concordance, un tableau de décomptes), et plus précisément penser et outiller [3] l'analyse comme un enchaînement d'étapes simples où la progression est guidée par la visualisation de chaque état intermédiaire ; (3) il faut rendre visibles aussi bien tous les paramètres des requêtes que l'impact qu'ils ont sur les données obtenues, pour guider efficacement l'interprétation.

En effet, l'utilisateur visé est familier de ses données. C'est à partir de leur consultation, de leur exploration, qu'il peut souhaiter en construire diverses vues synthétiques ou déployées, extraites de leur contexte ou (re-)contextualisées. L'important est qu'ici, en toutes circonstances, pour notre utilisateur, le texte est premier, c'est la référence et le point d'appui.

[2] Initialement développé en 2007-2008 dans le cadre d'une délégation CNRS de MH Lay, autour du projet ANR Textométrie (http://textometrie.ens-lsh.fr). Les financements des développements informatiques sont le fruit d'une collaboration avec le CESR de Tours, développeur informatique : François Raynaut.

[3] On outille la conception de l'analyse comme succession d'étapes notamment en documentant automatiquement chaque état par un rappel des différentes opérations qui ont conduit à lui, avec la possibilité donc de revenir sur un état antérieur de l'analyse pour réexaminer les choix faits.

L'accès aux outils ne se fait donc pas par la compréhension d'un environnement de travail organisant toute une série de fonctionnalités qui, dans l'absolu (abstraitement), peuvent s'enchaîner (plus ou moins facilement) les unes aux autres pour répondre à divers besoins. Les formulations abstraites de tri ou de sélection, dès lors qu'elles sont coupées de leur environnement textuel, ne font pas particulièrement sens pour un tel utilisateur ; même la notion de pourcentages de lemmes pose de difficiles questions, en raison de la multiplicité des bases de référence possibles (décompte sur les occurrences ou sur les types). Dans AnaLog, les opérations sont construites depuis les données visuellement disponibles : l'enchaînement des requêtes, des vérifications (ex: distributions et pourcentages) naissent de l'observation du résultat antérieurement fourni, et d'une action sur le tableau de données présenté.

Ainsi, consultant le résultat d'une concordance, l'utilisateur peut vouloir retourner au texte [4], ou inversement explorer plus finement le sous-corpus associé. Par exemple, après avoir identifié les contextes des mots en **ment*, on peut vouloir sélectionner ceux où le mot en *ment* est précédé d'un verbe. L'analyse n'est pas vécue comme la mise au point d'une requête complexe, on précisée peu à peu pour être relancée sur le texte initial et en tirer directement une concordance "définitive" [5], AnaLog propose une conception de l'analyse en étapes élémentaires, plus proche du raisonnement des utilisateurs concernés, et permettant plus facilement de revenir méthodiquement sur la construction de l'analyse. En effet, on rend compte ainsi d'un raisonnement qui progresse par étapes et qui est guidé par l'observation des états successifs obtenus ; alors que revenir toujours à l'état initial et intégrer tous les éléments de l'analyse en une seule requête est perçu comme artificiel et complexe, voire dénué de sens.

1.3. Croisement syntagmatique – paradigmatique : pouvoir observer et retravailler le tissage du texte avec les données issues des ressources d'annotation

L'accès aux corpus a radicalement évolué au cours des deux dernières décennies. En particulier, là où les bibliothèques numériques ne proposaient encore que des fichiers bruts, de type image [6] ou texte [7], on voit se multiplier les ressources annotées, entre autres au niveau linguistique [8]. On a des initiatives tant au niveau de l'édition de l'œuvre d'un auteur [9], qu'au niveau de la diffusion de collections nationales [10].

Disposer de corpus annotés permet d'élargir considérablement les possibilités de requêtes, et d'approfondir l'observation des données, par le biais même des métadonnées. AnaLog propose d'exploiter de façon particulièrement étendue ces nouvelles ressources, en interrogeant non seulement les corpus, observables construits, mais aussi, simultanément, leurs modèles de description. De fait, le corpus annoté est le fruit de plusieurs sources d'informations qui se trouvent tissées en un point donné : le texte et les différentes métadonnées dont il a été enrichi, ces métadonnées pouvant être décrites comme un tout cohérent disponible par ailleurs, un modèle de description. Pour donner du sens aux analyses portant sur les métadonnées, il est utile de pouvoir faire porter les requêtes sur le texte annoté (usage des métadonnées en contexte), mais aussi sur les valeurs d'étiquette disponibles qui auraient pu être retenues au moment de

[4] Cette fonctionnalité de retour au texte est déjà traditionnellement présente dans les logiciels de textométrie.
[5] Pour l'exemple donné, il suffirait d'une requête demandant tout de suite les verbes suivis d'un mot en *-ment*.
[6] Par exemple Gallica 1 : http://gallica.bnf.fr/.
[7] Par exemple les versions initiales de la base Epistemon : http://www.bvh.univ-tours.fr/Epistemon/.
[8] Ces textes annotés suivent les recommandations de la TEI-P5 (Burnard, 1995), http://www.c-tei.org.
[9] Par exemple les manuscrits de Stendhal (Lebarbé, 2009), http://manuscrits-de-stendhal.org.
[10] Par exemple, les travaux menés à Berlin pour les archives allemandes, http://www.deutschestextarchiv.de.

l'annotation (avant la désambiguïsation). En effet, les choix faits dépendent des ressources en présence comme de l'intention présidant à l'opération d'étiquetage, et il est essentiel de pouvoir les mettre en perspective. Autrement dit, contextualiser un élément annoté ne se limite pas à l'observation de l'ordre syntagmatique dans lequel il est immergé, on doit aussi pouvoir le situer dans l'ordre paradigmatique auquel il appartient. Une requête de concordance peut porter tant sur les occurrences en contexte que sur les ressources paradigmatiques, pour les valeurs actualisées (texte annoté) ou latentes (autres valeurs possibles suivant la ressource).

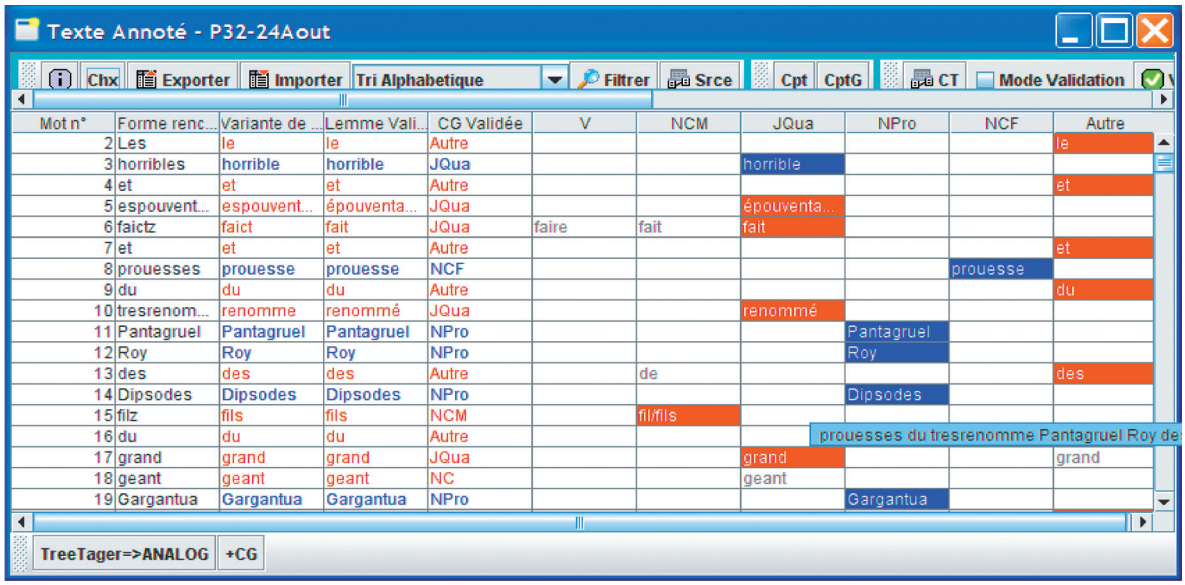
L'utilisateur humaniste est particulièrement sensible à ce travail conjoint sur les textes et sur les ressources descriptives : il permet la prise en compte de la relativité des descriptions, l'adaptation d'une ressource d'annotation donnée à un texte toujours singulier. AnaLog a ainsi été appliqué à la comparaison de textes (deux éditions de Pantagruel) ou d'annotation (même texte de Rabelais étiqueté selon deux dictionnaires).

2. Présentation du logiciel

2.1. La représentation conjointe du texte et d'une ressource de description

2.1.1. Vue générale

AnaLog considère qu'un corpus annoté est le tissage entre deux dimensions, l'une syntagmatique, le texte, l'autre paradigmatique, une ressource descriptive (un lexique, une ontologie...) ; et au croisement de ces deux dimensions se trouve l'annotation réalisée sur le texte au moyen de la ressource. Visuellement l'interface (Fig. 1) met donc en scène les trois espaces, et on peut utiliser les outils de requête et de modification des ressources sur les trois zones.



Mot n°	Forme renc...	Variante de ...	Lemme Vali...	CG Validée	V	NCM	JQua	NPro	NCF	Autre
2	Les	le	le	Autre						le
3	horribles	horrible	horrible	JQua			horrible			
4	et	et	et	Autre						et
5	espouvent...	espouvent...	épouventa...	JQua			épouventa...			
6	faictz	faict	fait	JQua	faire	fait	fait			
7	et	et	et	Autre						et
8	prouesses	prouesse	prouesse	NCF					prouesse	
9	du	du	du	Autre						du
10	tresrenom...	renomme	renommé	JQua			renommé			
11	Pantagruel	Pantagruel	Pantagruel	NPro				Pantagruel		
12	Roy	Roy	Roy	NPro				Roy		
13	des	des	des	Autre		de				des
14	Dipsodes	Dipsodes	Dipsodes	NPro				Dipsodes		
15	filz	fils	fils	NCM		fil/fils				
16	du	du	du	Autre			prouesses du tresrenomme Pantagruel Roy de			
17	grand	grand	grand	JQua			grand			grand
18	geant	geant	geant	NC			geant			
19	Gargantua	Gargantua	Gargantua	NPro				Gargantua		

Figure 1 : Affichage conjoint du texte et de la ressource

Ce tableau (Fig. 1) se structure donc en trois zones :

Zone du texte	*Zone de l'annotation*	*Zone de la ressource descriptive*
2 colonnes, écrites en noir	Ici 3 colonnes, en fait autant de colonnes que d'informations apportées par l'annotation	Colonnes suivantes, en fait autant de colonnes que de valeurs pour la (ou les) propriété(s) considérée(s).

2.1.2. Zone du texte brut

La zone représentant la dimension syntagmatique, à savoir le texte dans son déroulement, occupe les deux premières colonnes, à gauche. La première colonne (*Mot n°*) indique le numéro d'ordre des mots dans le texte. La deuxième colonne (*Forme rencontrée*) permet de lire le texte verticalement. Un tri sur la première colonne permet de restituer une lecture linéaire : la séquence des lignes est en effet systématiquement modifiée par toute opération de tri ou de filtrage.

2.1.3. Zone de l'annotation du texte

Les colonnes de cette zone médiane présentent les annotations retenues pour le texte. On y trouve, pour chaque occurrence et pour chaque propriété, la valeur de la ressource affectée à l'occurrence. Dans l'exemple, les informations portées par l'annotation sont : la *Variante de lemme* (lemme correspondant à la forme trouvée dans le texte, avant réduction des variantes), le lemme (lemme normalisé), la catégorie grammaticale, un descripteur sémantique, et le mode de validation de l'annotation.

2.1.4. Zone de la ressource descriptive

La troisième et dernière zone correspond aux informations en provenance de la dimension paradigmatique. Elles sont générées par la lecture de la ressource. Les têtes de colonne correspondent aux étiquettes disponibles dans la ressource pour la (ou les) propriété(s) considérée(s), ici le jeu d'étiquettes pour la catégorie grammaticale. Les valeurs apparaissant dans les cases présentent les informations complémentaires alors présentes dans la ressource pour l'étiquette considérée, ici le(s) lemme(s) associable à la forme. Dans une tâche de validation, on trouve donc dans cette zone, de façon organisée, les ressources potentiellement disponibles pour la description de chaque occurrence.

2.1.5. Usage de ce tableau et des différentes zones

Les trois espaces peuvent être étudiés ensemble ou séparément. On peut se concentrer sur l'analyse du texte annoté (désambiguïsé), mais on peut aussi vouloir observer la simple association du texte brut aux annotations possibles, analyser le texte brut seul, l'ensemble des données disponibles seules, ou encore les trois espaces conjointement.

L'analyse du texte brut ou celle du texte annoté se pratiquent déjà couramment en textométrie. La possibilité d'y associer l'analyse et l'exploration des ressources d'annotation est plus originale, et ouvre d'autres types d'investigations, comme l'observation des lacunes, redondances, zones d'ambiguïté dans le modèle de description. La présentation visuelle de toutes les informations disponibles, et la possibilité de les présenter « en empilement » (par une concordance, Pincemin et al., 2006) sur des contextes syntagmatiques permet d'observer la systématicité des descriptions et de repérer les types d'erreur le cas échéant.

2.2. Types de traitements

L'exploitation de grands corpus montre vite les limites d'une lecture linéaire. Opérer des rapprochements entre occurrences distantes fait apparaître des redondances :

- l'observation de toutes les occurrences d'un phénomène dans un certain contexte permet de formuler une hypothèse (éventuellement généralisable) ;
- le traitement dans le même temps de toutes ces occurrences permet de garantir une certaine stabilité de l'analyse dans le contexte syntagmatique et paradigmatique considéré, prenant en compte les propriétés du type pour décrire l'individu.

Ceci a donc conduit à développer un outil de filtrage généralisé, qui peut se déployer sur toutes les colonnes d'un tableau de données. Si l'on considère le tableau représentant le texte annoté et la ressource utilisée, le filtre peut donc s'appliquer (1) au niveau des formes rencontrées dans le texte, (2) au niveau des annotations retenues, comme (3) au niveau des ressources disponibles. Par ailleurs cet outil est associé à un « index généralisé » qui permet de faire des dénombrements et des calculs de pourcentages sur les résultats de concordance.

On voit ici le parti pris des outils développés dans AnaLog : on met largement l'accent sur le couplage entre annotation (procès et résultat) et fonctionnalités de tri, de sélection, concordances, fonctionnalités ici repensées et déclinées ; par contre, on n'a pas recours à des statistiques avancées, peu familières aux utilisateurs visés : on se limite à des opérations de dénombrement et de calculs de pourcentages.

2.3. L'expression de requêtes à partir des données

Une requête s'exprime comme un filtre sur un tableau de données. Basée sur la structure du tableau, elle donne ainsi à formuler la requête sous une forme proche des données observées comme du résultat recherché, forme plus intuitive pour l'utilisateur humaniste qu'une équation dans un langage formel de requête.

2.3.1. La requête peut mobiliser plusieurs propriétés, de toutes natures

Le formulaire de requête reprend la structure du tableau central de l'application. Chaque colonne est vue comme une propriété sur laquelle on peut poser (ou pas) une condition de sélection (par une expression régulière). On peut ainsi formuler une question comme :

[SELECTION 1] : « afficher -en les entourant d'un contexte de 5 mots à gauche et 5 mots à droite- la liste de formes qui ont été validées *Verbe* et qui étaient par ailleurs des adjectifs qualificatifs dans la ressource d'annotation » (Fig. 2).

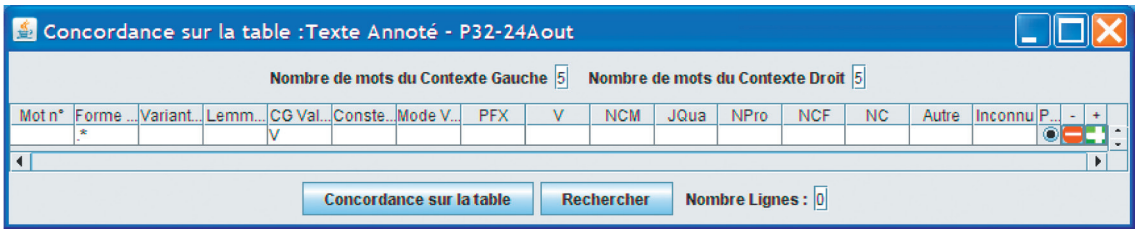


Figure 2 : Formulaire de requête de concordance pour la [selection 1]

Par défaut la réponse reprend la totalité des colonnes du tableau de départ. Mais on peut choisir de n'afficher que les colonnes jugées pertinentes (Fig. 3).

On peut voir ici une autre façon de poser une question très proche en apparence, mais qui donne pourtant à voir des choses bien différentes sur les données, cette deuxième requête permettant de s'interroger sur la cohérence d'ensemble du processus de validation :

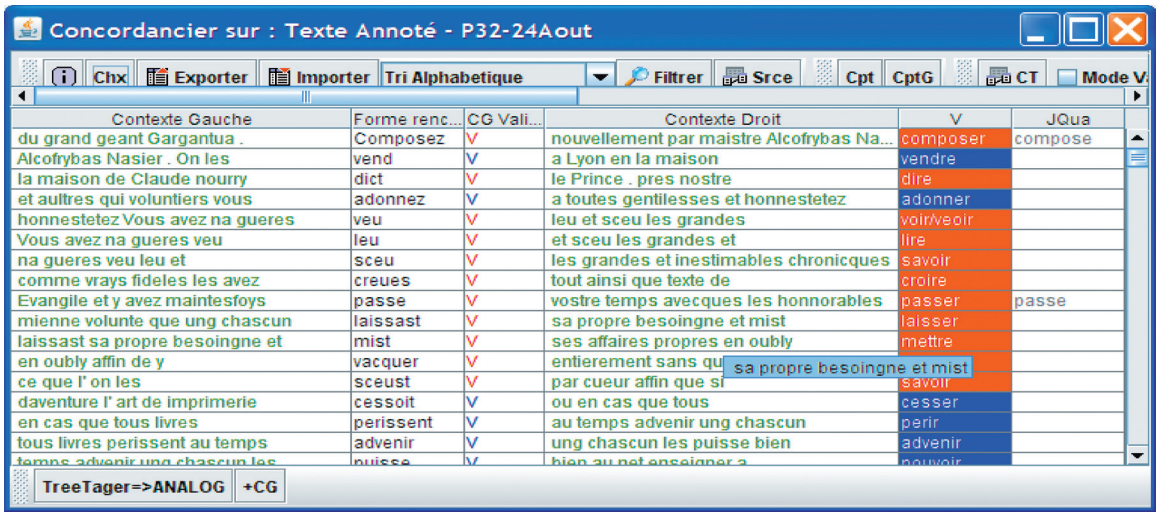


Figure 3 : Affichage de la concordance pour la [selection 1]

[SELECTION 2] : « indépendamment de la catégorie finalement retenue, afficher les formes en contexte en montrant celles décrites comme *Verbe* ou *Adjectif qualificatif* » (Fig. 4).

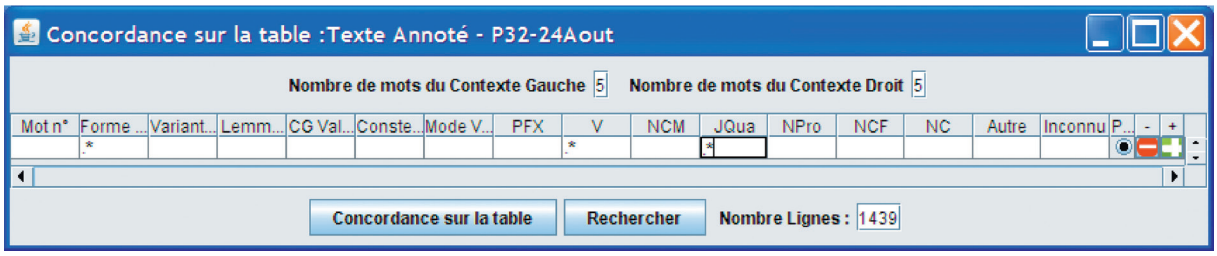


Figure 4 : Formulaire de requête de concordance pour la [selection 2]

Une telle requête peut mettre en évidence les difficultés rencontrées à choisir entre un participe et un adjectif, peut aussi mettre en évidence des régularités permettant d'homogénéiser les descriptions (les choix validés sont surlignés en rouge) (Fig. 5).

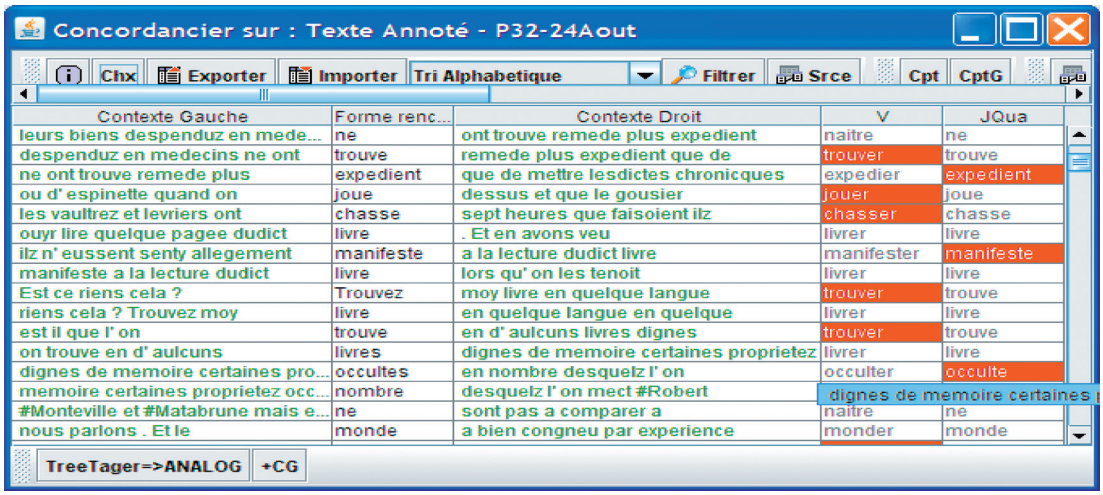


Figure 5 : Affichage de la concordance pour la [selection 2]

2.3.2. La requête peut décrire un motif s'étendant sur plusieurs positions : le pivot articulé

On peut faire porter le pivot sur plusieurs positions d'occurrences successives, donc sur une suite de plusieurs lignes. Au niveau du formulaire de requête, cela revient tout simplement à ajouter autant de lignes que nécessaire pour décrire le motif. Comme on peut décrire chacun des éléments du pivot par une ou plusieurs propriétés (sur chaque ligne), on peut ainsi rechercher des motifs de faisceaux de propriétés [11].

Dans le cas suivant, on veut observer la structure *Verbe + de + mot grammatical*, et on demande à mettre en évidence également le mot qui précède et le mot qui suit, avec la possibilité de trier facilement sur eux (Fig. 6).

La restitution du résultat de la requête permet de visualiser le contexte du pivot articulé. On peut alors affiner les observations par des opérations de tri : chacune des colonnes peut en effet être triée en tant que telle. Un tri alphabétique sur la colonne de mots grammaticaux suivant la préposition *de* permettra par exemple de se faire une idée sur la distribution des déterminants dans cette position (Fig. 7).

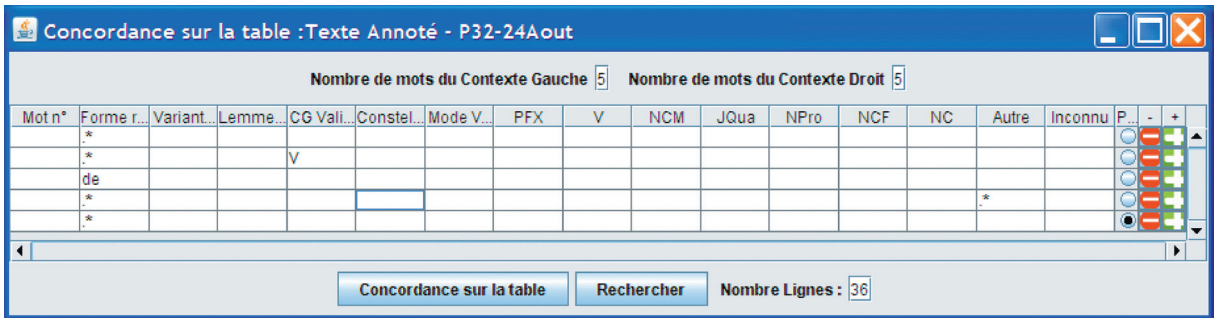


Figure 6 : Exemple de pivot articulé

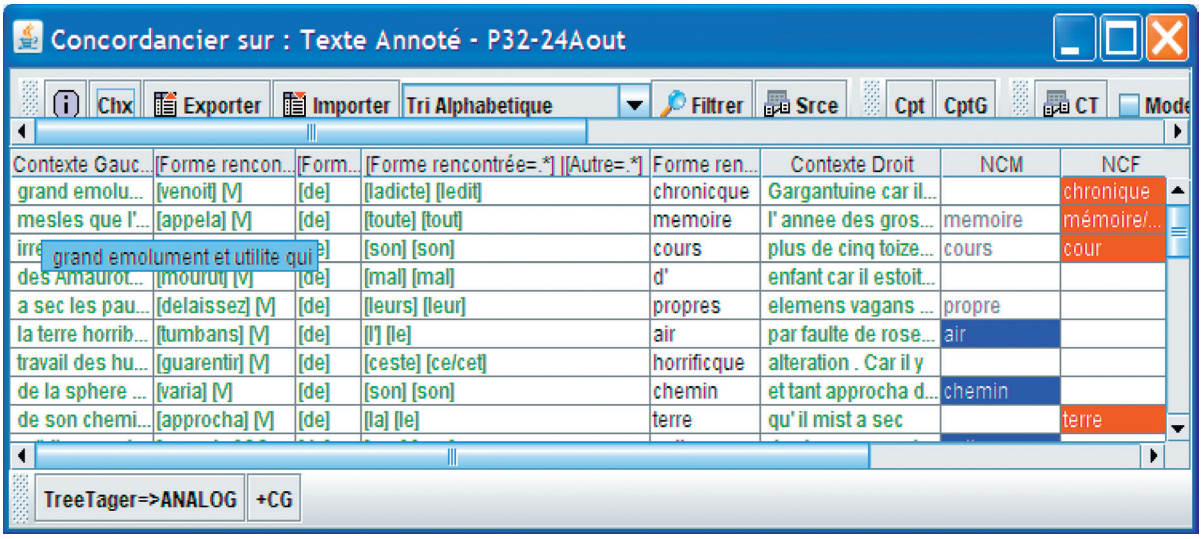


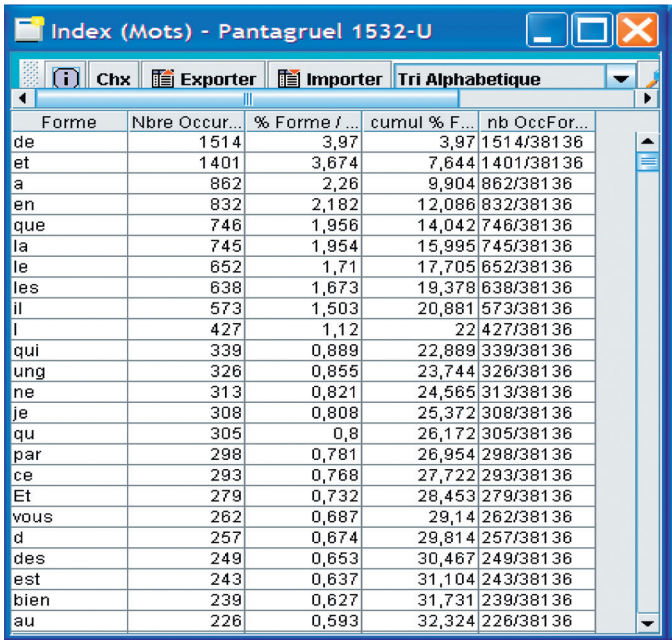
Figure 7 : Exemple de résultat de concordance sur pivot articulé

[11] D'autres outils d'interrogation offrent la possibilité de rechercher des motifs de faisceaux de propriétés, tels que CQP (Schulze, 1994, Christ, 1994, Schulze and Christ, 1996) utilisé dans le logiciel de textométrie Weblex (Heiden, 2002), ou encore ConcQuest (Kraif, 2008). Le principal apport d'AnaLog tient au fait de formuler les requêtes à partir de la visualisation des données, et de travailler sur les trois zones présentées au §2.1 (texte, annotation, ressource).

2.4. Dénombrements

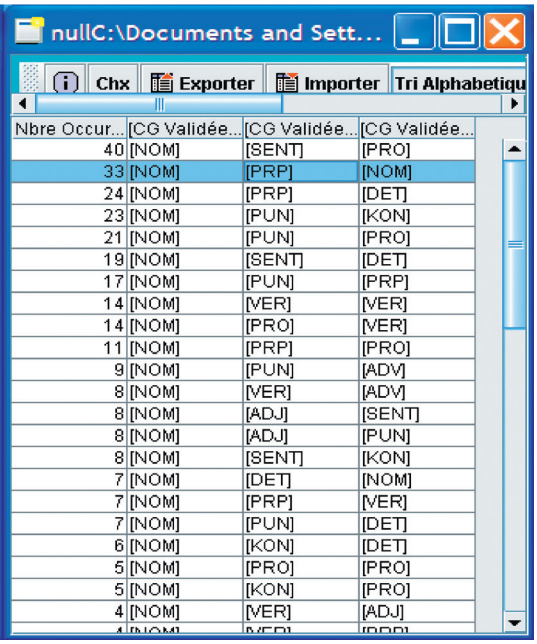
La concordance présente toutes les occurrences du pivot, qu'elle contextualise. La fonctionnalité Index [12] d'AnaLog fait quant à elle l'inventaire hors contexte des réalisations du filtre. Dès lors, on passe de l'occurrence au type, type pour lequel on peut donner des informations de dénombrement (nombre d'occurrences) et de pourcentages (dans l'espace sélectionné, les occurrences d'un même type couvrent *n*% de l'ensemble des occurrences) (Fig. 8).

Cette opération de comptage groupé est généralisée dans AnaLog à tous les tableaux que l'on trouve : le tableau central lui-même, par sélection de colonnes, les tableaux de résultats obtenus par requête, qu'elle porte sur le texte, le texte annoté ou les informations d'annotation disponibles. On peut ainsi construire des indicateurs quantitatifs sur les occurrences de toutes les structures observées : par exemple, rechercher, pour un texte, les suites de trois catégories grammaticales les plus représentées (Fig. 9).



Forme	Nbre Occur...	% Forme / ...	cumul % F...	nb OccFor...
de	1514	3,97	3,97	1514/38136
et	1401	3,674	7,644	1401/38136
a	862	2,26	9,904	862/38136
en	832	2,182	12,086	832/38136
que	746	1,956	14,042	746/38136
la	745	1,954	15,995	745/38136
le	652	1,71	17,705	652/38136
les	638	1,673	19,378	638/38136
il	573	1,503	20,881	573/38136
l	427	1,12	22	427/38136
qui	339	0,889	22,889	339/38136
ung	326	0,855	23,744	326/38136
ne	313	0,821	24,565	313/38136
je	308	0,808	25,372	308/38136
qu	305	0,8	26,172	305/38136
par	298	0,781	26,954	298/38136
ce	293	0,768	27,722	293/38136
Et	279	0,732	28,453	279/38136
vous	262	0,687	29,14	262/38136
d	257	0,674	29,814	257/38136
des	249	0,653	30,467	249/38136
est	243	0,637	31,104	243/38136
bien	239	0,627	31,731	239/38136
au	226	0,593	32,324	226/38136

Figure 8 : Index de tous les mots du corpus (formes graphiques)



Nbre Occur...	[CG Validée...	[CG Validée...	[CG Validée...
40	[NOM]	[SENT]	[PRO]
33	[NOM]	[PRP]	[NOM]
24	[NOM]	[PRP]	[DET]
23	[NOM]	[PUN]	[KON]
21	[NOM]	[PUN]	[PRO]
19	[NOM]	[SENT]	[DET]
17	[NOM]	[PUN]	[PRP]
14	[NOM]	[VER]	[VER]
14	[NOM]	[PRO]	[VER]
11	[NOM]	[PRP]	[PRO]
9	[NOM]	[PUN]	[ADV]
8	[NOM]	[VER]	[ADV]
8	[NOM]	[ADJ]	[SENT]
8	[NOM]	[ADJ]	[PUN]
8	[NOM]	[SENT]	[KON]
7	[NOM]	[DET]	[NOM]
7	[NOM]	[PRP]	[VER]
7	[NOM]	[PUN]	[DET]
6	[NOM]	[KON]	[DET]
5	[NOM]	[PRO]	[PRO]
5	[NOM]	[KON]	[PRO]
4	[NOM]	[VER]	[ADJ]

Figure 9 : Index des suites de trois catégories

La concordance et l'index apparaissent donc en pratique comme deux points de vue complémentaires possibles sur un même résultat, le premier (concordance) déployé (Fig. 10), le second (index) synthétique (Fig. 11). Ici il s'agit de la requête « VERBE + *de* ».

L'espace sur lequel porte le calcul est toujours défini par le tableau considéré : sa taille est le nombre total de lignes. Dénombrement des occurrences correspondant à un type et pourcentages sont donc liés au nombre de lignes du tableau considéré correspondant à ce type (et donc pas nécessairement le nombre d'occurrences en corpus).

[12] Dans la typologie des fonctionnalités textométriques présentée par ailleurs dans cette même conférence, l'Index d'AnaLog relève de la méta-fonctionnalité Vocabulaire.

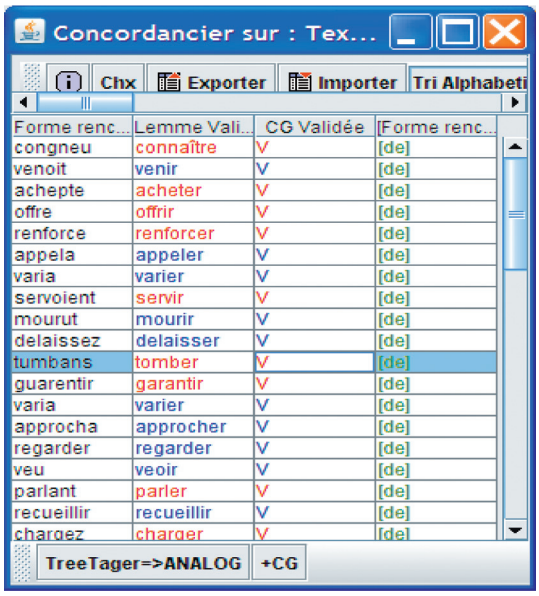


Figure 10 : Vue déployée du résultat d'un filtre, par une concordance

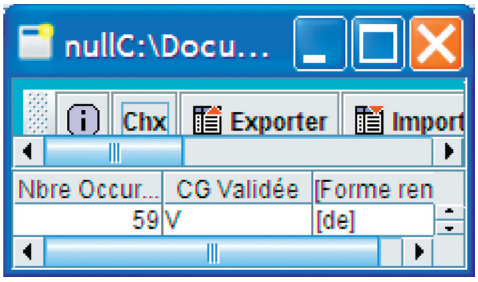


Figure 11 : Vue synthétique du résultat d'un filtre, par un index

2.5. Lien entre requête, affichage et décomptes

Dans la requête, toute colonne sur laquelle on pose un filtre, fût-il nul (joker .*, réalisé quelle que soit la valeur rencontrée), est affichée au niveau du tableau résultat. Cela est particulièrement sensible pour les pivots articulés, où seul l'un des éléments du motif (choisi librement par l'utilisateur) garde par défaut toutes les informations d'annotation et de valeurs latentes dans la ressource ; pour les autres éléments, pour garder telle ou telle information, il faut l'avoir activée par l'ajout d'un filtre. Or, ensuite, les décomptes se basent sur les valeurs affichées dans les colonnes sélectionnées. Donc la manière d'exprimer la requête permet de préparer le type de décomptes réalisés.

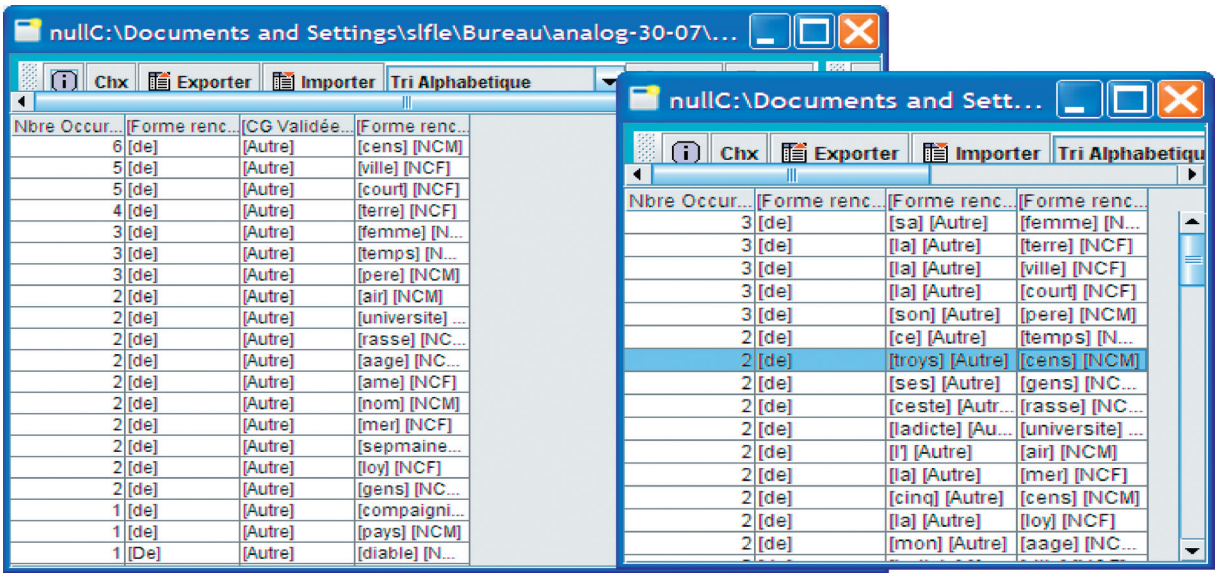


Figure 12 : Exemple de variation de la finesse de description (à partir de l'expression de la requête)

Ainsi, les possibilités d'expressions de l'équation de recherche permettent d'observer progressivement une donnée : par exemple ici (Fig. 12) la répartition des déterminants après la préposition. Dans la concordance, on a recherché les pivots structurés de la forme « préposition *de* suivie d'une catégorie grammaticale mineure suivi d'un nom commun ». Dans la copie d'écran de gauche, le degré de finesse de la description est celui qui vient d'être donné, dans celle de droite, la forme graphique prise par le déterminant est précisée.

3. Conclusion : les principales innovations d'AnaLog et leur intérêt pour d'autres contextes

En proposant une réponse logicielle aux besoins et aux manières de travailler de chercheurs, spécialistes de textes mais peu enclins à l'usage de statistiques, AnaLog reprend des principes et des fonctionnalités centrales de la textométrie (le retour au texte, les concordances, l'index hiérarchique), tout en renouvelant plusieurs aspects touchant à l'annotation des corpus et à une ergonomie en prise directe sur les données.

L'exemple de réalisation présenté considère l'enrichissement du corpus par une source de type dictionnairique, avec des informations essentiellement morphologiques ; mais la structure de données est transposable à des informations d'autres natures et s'adapte à d'autres domaines d'analyse. L'intérêt est de contextualiser chaque unité non seulement dans la séquence syntagmatique et l'environnement textuel, non seulement par son positionnement dans une partition structurant le corpus, mais aussi par ses potentialités de description dans la ressource qui sert de référentiel pour l'étiquetage, afin de mieux comprendre et contrôler l'annotation.

L'interface expérimente une généralisation de la présentation en tableau pour tous types de données (syntagmatiques et paradigmatiques, synthétiques ou déployées), pour proposer des outils simples et unifiés de traitement sur ces tableaux (tris, filtrages, décomptes groupés). L'ergonomie passe aussi par une explicitation et une visualisation systématique des étapes et des choix d'analyse.

Une des voies de développement et de renouvellement de la textométrie reste bien celle qui allie l'attention très concrète au mode de travail et de pensée d'une communauté d'utilisateurs, et la création de nouveaux outils logiciels.

References

Antoni-Lay M.H. and Demonet M.L. (2000). Adaptation d'un lemmatiseur au corpus rabelaisien : naissance d'Humanistica. In *Actes des JADT2000*, Ecole Polythechnique Lausanne.

Brunet E. (2001). Le Logiciel Hyperbase. *L'Astrolabe.* http://www.uottawa.ca/academic/arts/astrolabe/auteurs.htm.

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

Christ O. (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* (CMP-LG archive id 9408005), Budapest, pp. 23-32.

Demonet M.L. (1998). Pronostiquer avec Hyperbase. In *Mots chiffrés et déchiffrés. Mélanges offerts à Etienne Brunet*, Genève : Slatkine, pp. 455-471.

Demonet M.L. and Lay M.H. (2008). Digitizing European Renaissance prints: a 3-year experiment on image-and-text retrieval, Kolkata. In *International Workshop on Digital Preservation of Heritage (IWDPH07)*.

Heiden S. (2002). *Weblex. Manuel Utilisateur. Version 4.1*. Laboratoire ICAR, UMR 5191, ENS Lyon.

Heiden S. (2006). Un modèle de données pour la textométrie : contribution à une interopérabilité entre outils. In *JADT2006.* Presses Universitaires de Franche-Comté, Besançon, pp. 487-498.

Kraif O. (2008). Comment allier la puissance du TAL et la simplicité d'utilisation? L'exemple du concordancier bilingue ConcQuest. In *JADT2008*, pp. 625-633.

Lebarbé T. (2009). Du corpus littéraire au corpus linguistique : dématérialisation, restructuration, lectures rhizomatiques et analyse linguistique des manuscrits. In Viprey, J-M. and Adam, J.-M., editors, *Corpus*, 8.

Lebart L. and Salem A. (1994). *Statistique textuelle*. Paris : Dunod.

Loiseau S. (2007). CorpusReader : un dispositif de codage pour articuler une pluralité d'interprétations. *Corpus*, 6: 153-186.

Muller C. (1977). *Principes et méthodes de statistique lexicale.* Paris : Hachette université.

Pincemin B., Issac F., Chanove M. and Mathieu-Colas M. (2006). Concordanciers : thème et variations. In *JADT2006*, Besançon : Presses Universitaires de Franche-Comté, pp. 773-785.

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

Schulze B.M. (1994). Entwurf und Implementierung eines Anfragesystems für Textcorpora. *Diplomarbeit Nr. 1059.* Universität Stuttgart, Institut für maschinelle Sprachverarbeitung (IMS) and Institut für Informatik.

Schulze B.M. and Christ O. (1996). The CQP User's Manual, Version 1.6. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/.