in *Sergio Bolasco, Isabella Chiari, Luca Giuliano (eds)*, Statistical Analysis of Textual Data -Proceedings of 10th International Conference JADT 2010, *Edizioni Universitarie di Lettere Economia Diritto, Roma, June 9-11, 2010. Available on-line in the* LED electronic archive *at the* `http://www.ledonline.it` *web site.*

Fonctionnalités textométriques : Proposition de typologie selon un point de vue utilisateur

Bénédicte Pincemin[1], Serge Heiden[2], Marie-Hélène Lay[3], Jean-Marc Leblanc[4], Jean-Marie Viprey[5]

[1]CNRS et [1,2]Université de Lyon – ICAR UMR 5191–
ENS-LSH – 15 parvis René Descartes – B.P.7000 – F69342 Lyon cedex 07 – France

[3]Université de Poitiers – Laboratoire FoReLL –
MSHS de Poitiers – 95 avenue du Recteur Pineau – F86000 Poitiers – France

[4]Université Paris-Est Créteil Val de Marne – Céditec EA 3119 –
61 avenue du Général de Gaulle – F94000Créteil – France

[5]Université de Franche-Comté et MSH de Franche-Comté – Archives, Textes, Sciences des Textes EA 3187 – 30 rue Mégevand – F25030 Besançon – France

Abstract

In the field of textometry, every text analysis software offers its own set of functionalities. Although these functionalities originate in common ideas, their implementations and their names may differ. It is therefore impossible to compile a comprehensive and synthetic list of functionalities straight, or to draw a direct comparison between softwares.
This paper sets a structured overview of textometric metafunctionalities. Each of them adresses a type of textual research: text reading and browsing (through different views), listing of units (typically words) with quantitative information, looking at the positions of units in the corpus, paradigmatic and syntagmatic linking of units, texts or properties. This functional typology covers the whole range of textometric processing, from context and parameters initialization to results management and analysis.
This typology was elaborated in order to design a new textometric software (project ANR-06-CORP-029). It may also be used to introduce to textometric functionalities, or to characterize software.

Résumé

Chaque logiciel de textométrie donne accès à un certain nombre de calculs. Mais en l'état actuel, l'inventaire synthétique des fonctionnalités, comme la comparaison des logiciels au plan fonctionnel, ne sont pas possibles directement, car, malgré des parentés d'approche, les calculs sont rarement strictement identiques et l'usage des noms désignant les fonctionnalités n'est pas uniforme.
Cette communication propose un recensement large et structuré des fonctionnalités de calcul textométriques en métafonctionnalités, correspondant chacune à un type de questionnement du corpus : lecture du texte (selon différents modes complémentaires), inventaires et décomptes d'unités (typiquement, de mots), étude de la position d'unités dans le corpus, associations syntagmatiques ou paradigmatiques d'unités, de textes, ou de propriétés. La typologie s'étend en amont des calculs aux fonctionnalités de mise en place de l'environnement d'interrogation, et en aval aux fonctionnalités de gestion et d'aide à l'analyse des résultats des calculs.
Cette typologie a été mise au point pour la conception d'une nouvelle plateforme textométrique (projet ANR-06-CORP-029). Elle peut également être utile pour présenter un panorama des possibilités d'analyse textométrique, ainsi que pour situer différents logiciels au plan fonctionnel.

Keywords : textometry, lexicometry, textual statistics, functionnalities, software design, textual analysis uses.

1. Contexte et positionnement

1.1. Problématique

Quels traitements offre la textométrie pour l'analyse textuelle d'un corpus numérisé ? Plusieurs situations peuvent conduire à établir un inventaire de fonctionnalités :

- une présentation pédagogique générale de la textométrie, dans un cours, un article de synthèse, une thèse, ou un ouvrage de référence comme (Lebart & Salem 1994) ;
- un exposé des possibilités de traitement offertes par un logiciel donné : manuels utilisateur, tels que ceux de Hyperbase (Brunet 2006), Lexico 3 (Fracchiolla & *al*. 2003), Le Trameur (Fleury 2009), Weblex (Heiden 2002)[1], et supports de formation ; l'interface d'un logiciel suppose aussi des choix de mise en scène des fonctionnalités, dans l'organisation des menus déroulants ou des boutons d'appel des fonctionnalités ;
- l'établissement d'une grille pour décrire l'état de l'art et situer différents logiciels[2] ;
- une étude de synthèse en vue de la conception d'un nouvel outil.

C'est de ce dernier point de vue que relève la présente communication. Sa particularité est d'être prospectif ; il s'agit de prendre du recul pour (i) comprendre la logique des développements théoriques jusqu'à présent, (ii) saisir les équivalences entre des fonctionnalités analogues et inversement repérer les différences significatives à savoir utiliser, et (iii) percevoir des lieux d'innovation.

1.2. Conventions

Convenons pour la suite que nous étudions un corpus divisé en *parties*, analysées en *unités*. Parties et unités sont des désignations simples pour des rôles que peuvent prendre toutes sortes de composantes du corpus : les parties peuvent correspondre aux textes, à des regroupements de textes, à des qualifications de l'expression (par ex. toutes les répliques de chacun des personnages dans une pièce de théâtre) ; les unités se situent communément au plan lexical, mais elles peuvent aussi être infra- ou supra-lexicales. Unités comme parties peuvent être considérées sous l'angle d'une quelconque de leurs *propriétés* de description (par exemple pour un mot sa graphie, son lemme, sa catégorie morphosyntaxique ; pour un texte, son année de publication, son auteur, son genre). Les valeurs de la propriété considérée définissent des *types*, les unités *occurrences* au fil du texte liées à un même type en sont alors autant d'attestations. Nous appellerons *focus* l'objet (unité, suite d'unités pouvant être défini par un patron plus ou moins complexe) sur lequel se centre la recherche ou le calcul, le cas échéant.

[1] Nous appuyons explicitement cette étude sur les logiciels qui sont à la fois *(i)* les plus représentatifs de la textométrie telle que décrite dans (Lebart & Salem 1994) et *(ii)* les plus diffusés dans la communauté scientifique. D'autres logiciels existent bien sûr dans la même famille, et ont aussi été considérés dans la réflexion, comme Astartex, Taltac, DTM-Vic, Alceste, Sato, Xaira, le Sphinx-Lexica.

[2] Dans le cadre du projet Textométrie, D. Mayaffre a coordonné l'édition mise à jour et augmentée d'un *Tableau synthétique des fonctionnalités des logiciels de textométrie* (2007). Sur site *Textopol*, la discussion *Tableau récapitulatif des fonctionnalités* cite deux réalisations : http://tinyurl.com/LogicielsTextometrie (Ph. Gambette) et http://www.cati.paris4.sorbonne.fr/centre/activites/colloques/synoptiq_11mars05.htm (F. Deconinck).

1.3. Etat de l'art

Les fonctionnalités textométriques sont souvent présentées de façon peu structurée, bien qu'elles puissent être nombreuses[3]. Elles sont généralement groupées en au moins deux grandes catégories : les fonctionnalités *documentaires* couvrent les diverses formes de consultation du corpus et d'accès au texte, notamment via la recherche des contextes d'un focus ; elles sont opposées aux fonctionnalités *statistiques*, qui font appel à une modélisation mathématique avancée (probabilités, statistiques, analyse des données, etc.), et qui génèrent des listes d'unités accompagnées de scores, des tableaux de similarité ou de coordonnées. On parle quelquefois de fonctionnalités d'*exploration* textuelle, pour désigner les premières vues données à l'utilisateur après l'intégration de son corpus (consultation du texte, dictionnaire des mots du corpus avec leur fréquence), et de fonctionnalités de *navigation* textométrique pour rendre compte notamment des dispositifs donnant une représentation globale du corpus ou de résultats archivés, avec des accès hypertextes à des vues détaillées (tels la carte des sections et le rapport de Lexico 3).

Les fonctionnalités permettant de définir un focus sont souvent introduites avec les fonctionnalités documentaires. La transversalité de ces procédures de construction de focus pour différents calculs n'est pas toujours mise en évidence, en particulier dans l'interface, où la manière de définir un focus peut prendre des formes dispersées et variables selon le contexte. De même, les fonctionnalités de manipulation des sorties (annotation des représentations graphiques, tri des listes) ou les possibilités de parcours hypertextuels sont la plupart du temps décrites avec les fonctionnalités de calcul concernées.

En ce qui concerne les fonctionnalités documentaires, la fonctionnalité d'affichage du texte n'est pas toujours explicitée ; et la distinction entre concordance et affichage de contextes non centrés sur le focus est peu ou pas explicitée : du point de vue technique cela peut effectivement être vu comme des variantes d'affichage d'une même procédure, alors que du point de vue des usages ces présentations ont des propriétés très différentes.

Le dictionnaire des unités du corpus, les mesures générales souvent associées (taille du corpus, loi de zipf, longueur des phrases, accroissement du vocabulaire…), ainsi que le calcul des segments répétés, ont un statut très variable selon les présentations. Ces fonctionnalités sont tantôt perçues comme faisant partie de l'ouverture de la base (car définissant des unités), ou présentées comme relevant d'une première catégorie de fonctionnalités dites d'exploration ; on les trouve également assimilées à des fonctionnalités documentaires (lorsque ce sont les fonctionnalités préliminaires), ou même classées dans les fonctionnalités statistiques (puisqu'il y a modélisation de distributions ou synthèse des contextes locaux).

Les fonctionnalités statistiques ne sont généralement organisées que par leur ordre, avec des enchaînements dont on explicite quelquefois la motivation. L'ordre global de présentation est variable, mais certains enchaînements sont souvent observés. La partie s'ouvre généralement sur les fonctionnalités de décompte (dictionnaire, zipf, etc.) si elles n'ont pas été abordées ailleurs. Puis les graphes ou histogrammes de répartition apparaissent comme une première procédure simple d'étude de la distribution d'un focus. Sont ensuite abordés tantôt le calcul des spécificités (qui donne une mesure statistique de la distribution), tantôt un calcul de cooccurrences (qui vient comme une autre manière de caractériser l'environnement, cette

[3] Les tableaux comparatifs détaillés comportent plusieurs dizaines d'entrées, les manuels comptent jusqu'à une quarantaine de sections traitant chacune d'une fonctionnalité.

fois-ci lexical, d'un focus). Spécificités et cooccurrences sont par ailleurs souvent considérées comme liées du fait qu'une technique de calcul des cooccurrents recourt à un calcul de spécificités. Le cœur des procédures statistiques est la constitution de tableaux (de contingence, de similarités) et leur traitement par analyse des données (analyses factorielles, analyse arborée, classifications…). Enfin, les présentations se terminent souvent par les fonctionnalités dites topologiques ou topographiques implémentées plus récemment (mais elles peuvent aussi être associées aux cooccurrences, voire être exposées dans les premières fonctionnalités, comme point d'accès au texte). Viennent également dans les dernières les fonctionnalités spécifiques à certains types de corpus comme les corpus alignés.

La structure adoptée par l'ouvrage de référence (Lebart & Salem 1994) est relativement originale. Un chapitre consacré aux « unités de la statistique textuelle » concentre un grand nombre de fonctionnalités, vues donc sous l'angle de leur participation à la définition et à la caractérisation des unités d'analyse : non seulement dictionnaire, mesures générales et segments répétés, mais aussi concordances et cooccurrences, mettant au jour des régularités locales. Les chapitres suivants présentent des procédures d'analyse des données ; les spécificités sont introduites comme un moyen de caractérisation d'une partie d'un corpus, pour l'aide à l'interprétation d'une analyse factorielle ou d'une classification.

Enfin, en marge des fonctionnalités d'analyse à proprement parler, sont présentées les indispensables fonctionnalités d'import de corpus et d'ouverture d'une base déjà existante, ainsi que les fonctionnalités de sauvegarde, d'impression, et d'export dans des formats exploitables par d'autres logiciels. On peut également évoquer la gestion des traitements lourds, les données textuelles soumises à la textométrie étant souvent volumineuses.

1.4. Point de vue utilisateur

L'objectif de la présente typologie est d'adopter un point de vue utilisateur. Elle est structurée selon les différents types de besoins, de questionnements, que l'utilisateur peut mobiliser dans son étude du corpus. Ce point de vue s'écarte parfois d'un point de vue technique, qui peut considérer comme une même fonctionnalité ce qui repose sur le même calcul sous-jacent. Par exemple, le calcul des spécificités peut se retrouver pour nous dans au moins deux métafonctionnalités, la *Distribution* et les *Cooccurrences*. Et inversement, la *Distribution* se trouve réalisable dans différents calculs, d'un simple histogramme de répartition montrant les fréquences dans chaque partie, au calcul statistique des spécificités, comme à d'autres indices statistiques utilisés eux aussi pour évaluer une irrégularité de répartition.

1.5. Organisation et limites de l'exposé

Bien que la typologie proposée soit large, elle ne couvre pas des fonctionnalités essentielles mais non spécifiques à la textométrie, comme l'aide en ligne. Elle rend peu compte de l'ergonomie et des facilités de navigation, qui peuvent être très importantes notamment pour faciliter le retour au texte. La présentation de la typologie pourrait aussi donner lieu à une étude approfondie, fonctionnalité par fonctionnalité. On reporte à d'autres publications de tels développements. L'enjeu ici est de donner à comprendre une sorte de référentiel, en se limitant dans la description de chaque métafonctionnalité à ce qui permet de cerner sa portée et d'en reconnaître des implémentations.[4]

[4] Les parcours hypertextes peuvent alors ensuite être conçus pour lier les fonctionnalités, en fonction d'une méthodologie d'usage que l'on exprimerait dans les termes de la typologie.

Les entrées de la typologie sont en effet des métafonctionnalités : elles groupent sous une même problématique, différents traitements concourant à y répondre. Nous sommes conduits aussi à faire des propositions terminologiques pour désigner ces métafonctionnalités, mais ce sont plutôt les concepts sous-jacents qui se veulent le principal apport scientifique de l'article.

Tout d'abord (§2), une vue globale permet de percevoir la couverture et la logique d'ensemble de la typologie, avec ses principales articulations. Présentée sous la forme d'un tableau, elle fournit ainsi une fiche pratique, synthèse pour se remémorer la typologie, ou grille pour la description d'un logiciel. Puis (§3), nous détaillons la présentation des métafonctionnalités au cœur de la textométrie, celles correspondant aux calculs textométriques à proprement parler, notamment pour expliciter la logique de différenciation entre fonctionnalités traditionnellement proches voire confondues.

Par ailleurs, complémentairement, nous mettons en ligne une description de plusieurs logiciels textométriques selon cette typologie[5]. Son intérêt est d'abord de donner des exemples illustratifs d'implémentation des métafonctionnalités définis ici, aidant à mieux comprendre, concrètement, à partir de l'expérience d'un logiciel, en quoi peut consister telle ou telle métafonctionnalité. En pratique, c'est aussi une table de correspondance faisant le lien entre les fonctionnalités nommées dans le logiciel et les métafonctionnalités identifiées dans notre typologie, permettant ainsi d'accéder à divers logiciels avec une vue unifiée et orientée usages. Par exemple, à partir d'un besoin donné, comme celui d'étudier le positionnement des unités dans le déroulement linéaire du corpus, on obtient la ou les différentes fonctionnalités pouvant être mises en œuvre pour y répondre, dans les termes du logiciel considéré.

2. Vue globale de la typologie

Les entrées de la typologie sont des métafonctionnalités, sous lesquelles sont rassemblées des fonctionnalités de calcul et de traitement répondant à un même questionnement. Par exemple, la métafonctionnalité *Evolution* peut rassembler une fonctionnalité qui calcule la période caractéristique d'apparition d'une unité, et une autre fonctionnalité qui recherche les unités en progression globale et celles de plus en plus délaissées. On peut donc parler de la métafonctionnalité *Evolution*, lorsque l'on considère la problématique commune aux divers traitements rassemblés, et d'une fonctionnalité (de type) *Evolution* lorsque l'on a affaire aux différents traitements concrets.

La typologie s'ouvre sur les métafonctionnalités qui entrent naturellement en jeu au début d'une analyse textométrique, car elles établissent l'environnement dans lequel les calculs textométriques peuvent ensuite être lancés. Les deux grands groupes de métafonctionnalités suivants correspondent aux calculs textométriques à proprement parler, produisant des résultats que l'utilisateur doit interpréter. Le quatrième groupe vient donc logiquement compléter l'ensemble en couvrant les fonctionnalités qui outillent l'utilisateur dans ce travail interprétatif.

[5] Ce document est intitulé *Typologie des fonctionnalités textométriques selon un point de vue utilisateur : illustration par leurs implémentations dans des logiciels*, et est disponible sur le site du projet Textométrie (http://textometrie.ens-lsh.fr/), rubrique *Publications*. Cette partie de notre présentation est naturellement plus dynamique, car il s'agit de pouvoir suivre l'évolution des logiciels, comme d'étendre la description à d'autres logiciels que ceux considérés initialement.

Structuration	***Méta-fonctionnalités***	*Description brève*
DONNEES		Constitution de l'environnement à partir duquel lancer les calculs
Initialisation		
	Profil	Paramétrage par défaut, identification des traces…[6]
	Interfaçage	Importation : création et chargement d'une base textuelle[7] ; exportation de la base pour utilisation dans d'autres logiciels
	Session	Ouverture et fermeture d'une base textuelle, dans un état mémorisé et avec une archive des traitements précédents
Constructions		Définition de nouveaux objets, consultation et gestion d'objets réutilisables
	Propriétés	A partir des propriétés élémentaires disponibles, construction de nouvelles propriétés descriptives, par fusion de valeurs, croisement, etc.
	Sélections	Désignation d'un ensemble d'occurrences, en compréhension (moteur de recherche) ou en extension (par sélection sur une visualisation du corpus)
	Scénarios	Macro-traitement intégrant un enchaînement de traitements, pour l'automatiser
Vue courante		Paramètres généraux, exprimant un état de représentation des données
	Fond	Délimitation du sous-corpus étudié et détermination des unités prises en compte dans les calculs (et servant de référence aux calculs statistiques)
	Structure	Partition et regroupements actifs, le cas échéant ; parallélisme éventuel
	Localisations	Constitution et forme de l'indication précisant le positionnement d'une occurrence
	Plan d'analyse	Système d'unités considéré : segmentation, et propriété définissant les types
	Propriétés d'affichage	Propriété(s) utilisée(s) pour représenter les unités au niveau de l'affichage du texte ou d'un résultat de calcul.
	Focus	Unité ou motif sur lequel se centre le traitement, le cas échéant
LECTURE		Accès au texte, traitements textométriques affichant des occurrences
	Texte	Affichage du texte, en rendant compte visuellement de sa structure logique, avec possibilités de navigation
	Vue interne	Affichage de la représentation interne (structures, propriétés) correspondant au texte ou à un extrait
	Extraits	Liste triable d'extraits définis par un critère et localisés
	Concordance	Liste triable des occurrences d'un focus, alignées verticalement, entourées de leur contexte sur une seule ligne, et avec indication de leur localisation
SYNTHESES		Traitements textométriques affichant des types
Relevés		Décomptes
	Vocabulaire	Liste triable des unités avec leur fréquence (et leurs sous-fréquences si partition active), pour tout le sous-corpus ou correspondant à un focus
	Mesures	Caractérisations quantitatives du corpus (tailles, gamme des fréquences, indices évaluatifs synthétiques)
Positions		Caractérisation des localisations d'unités dans le corpus
	Déroulement	Disposition et régularité des positions d'unités au fil du texte.
	Distribution	Répartition contrastée d'unités dans les parties d'une partition
	Evolution	Répartition d'unités dans un corpus à partition chronologique ou sérielle

[6] C'est un profil d'usage plutôt qu'un profil utilisateur, car cela correspond en pratique à un contexte d'emploi : un même utilisateur peut avoir besoin de plusieurs profils, et un même profil peut être partagé par plusieurs utilisateurs. Un profil peut notamment servir à proposer un paramétrage par défaut adapté à un corpus.

[7] Cette fonctionnalité comprend les aspects concernant la segmentation en unités. Selon la manière de mettre en œuvre cette fonctionnalité, il est possible, ou non : (i) d'importer une segmentation complète encodée dans le corpus ; (ii) de proposer une segmentation interne, plus ou moins paramétrable, en l'absence de segmentation déjà présente dans le corpus ; (iii) de prendre en compte une segmentation partielle encodée dans le corpus et de proposer une segmentation complémentaire pour les passages non segmentés ; (iv) d'importer plusieurs segmentations alternatives.

Associations		Liens (syntagmatiques ou non) entre éléments de même nature : plusieurs unités, ou plusieurs parties, ou plusieurs propriétés.
	Séquences	Repérage d'enchaînements d'unités récurrents
	Cooccurrences	Repérage de la coprésence d'unités au sein d'un même contexte
	Analogies	Repérage de similarités entre unités ou parties, configuration du corpus (par analyse factorielle par exemple), classification
	Corrélations	Force et orientation du lien entre deux propriétés, caractérisées par leurs valeurs au fil des occurrences
ANALYSE		Conduite de l'analyse : qualification des résultats, progression des traitements
Gestion de la charge		Nécessaire car les calculs textométriques peuvent être complexes et les données et résultats volumineux
	Progression du calcul	Indicateur de progression du calcul (jauge), possibilité d'interruption (annulation)
	Volume des résultats	Information et possibilités de réduction (échantillonnages,…)
Visualisations		Choix d'une sémiotique de présentation et outils d'exploitation adaptés
	Tableaux	Transformations sur les lignes et les colonnes (déplacement, suppression, fusion), sur les valeurs (recodage, normalisation…)
	Graphiques	Courbe, diagramme en bâtons, histogramme, camembert… Aide à la lecture, styles d'édition,…
	Diagrammes	Graphes, dendogrammes ; aide à la lecture, styles d'édition…
	Cartes	Aide à la lecture, styles d'édition, projections…
Organisation		Heuristique de mise en forme des résultats accompagnant leur interprétation
	Filtrage, regroupements, tris	Possibilité de traduire visuellement une organisation des résultats au fur et à mesure de leur analyse
	Synopse et comparaison	Mise en regard de plusieurs résultats et outils d'aide à leur comparaison (projection, différence, tracé de relations…)
Annotation		
	Commentaire	Commentaire libre, signet ; non destiné à l'analyse par des calculs (non indexé pour cela)
	Edition du corpus	Modification du corpus (valeur d'une propriété, segmentation…)
Archive		
	Journal	Systématique – note les fonctions appelées et les paramètres – pour lecture et pour élaboration de scénarios
	Rapport	Edition sélective et documentée de résultats et export (dont impression)

3. Présentation plus détaillée des métafonctionnalités de calculs textométriques

3.1. Les fonctionnalités de Lecture

Ce premier groupe de fonctionnalités est fondamental car il outille ce qu'il est convenu d'appeler le « retour au texte » : tout résultat exprimé via des unités hors contexte doit s'interpréter en vérifiant les valeurs prises par ces unités en contexte, local (l'entour immédiat de l'unité, au plan syntagmatique) et global (par une indication de localisation). La délimitation des métafonctionnalités est ici essentiellement basée sur la disposition des informations présentées, plus que sur leur nature. En effet, la mise en page et les outils de manipulation et de parcours associés déterminent des usages appropriés.

3.1.1. Texte

La fonctionnalité *Texte* typique affiche le corpus dans son déroulement linéaire, en rendant compte visuellement de la structure logique des textes (paragraphes, vers, didascalies, sections, etc.). L'affichage peut être paginé ou continu. L'interface est dotée de possibilités de

navigation, par accès séquentiel ou accès direct. Un corpus parallèle aligné peut donner lieu à une présentation synoptique. Lorsqu'un focus est défini et activé, les occurrences correspondantes dans le texte sont mises en évidence, et un dispositif permet de naviguer d'occurrence en occurrence.

La fonctionnalité *Texte* instrumente un retour au texte perçu comme un feuilletage du texte. Elle est utilisée pour l'analyse textuelle d'un phénomène avec un contexte non limité en taille, et la perception du positionnement dans la structure textuelle. Elle peut convenir mieux que la métafonctionnalité *Extraits* à la visualisation des occurrences d'un focus très fréquent, ou concentré sur certaines zones du texte.

3.1.2. Vue interne

La métafonctionnalité *Vue interne* sert à comprendre comment l'application « voit » le texte, sur quelle représentation se fondent les traitements textométriques –la segmentation en occurrences, les valeurs des étiquettes-, pour contrôler l'interprétation des résultats. La *Vue interne* peut être complète ou partielle, limitée à un choix de propriétés.

Dans les autres métafonctionnalités de *Lecture*, divers procédés d'affichage de propriétés sont déjà possibles (cf. logiciel Xaira) : affichage d'escamots (bulles contextuelles *fly-over* ou *pop-up*), utilisation de couleurs pour rendre compte de catégories, etc. On peut aussi jouer sur les propriétés d'affichage. Néanmoins, la représentation des données internes peut être plus efficace dans d'autres formes de disposition que celles du *Texte*, de l'*Extrait* ou de la *Concordance* : typiquement en tableau (une ligne par occurrence, une colonne par propriété), voire en arborescence (pour la représentation de structures emboîtées).

3.1.3. Extraits

La métafonctionnalité *Extraits* fournit une série de passages, munis de références de localisation dans le corpus, et donnés dans un ordre choisi, avec des regroupements possibles. Le critère de sélection des extraits est de nature variable. Le cas de figure le plus courant est la recherche de contextes, telle qu'un relevé des paragraphes contenant tel focus. L'empan des contextes peut être défini par fenêtrage (nombre d'unités de part et d'autre du focus), délimiteur (typiquement ponctuation forte), ou structure englobante (le paragraphe, la réplique,…). Mais l'extrait n'est pas nécessairement un contexte autour d'un focus : on peut par exemple vouloir tous les premiers vers d'un recueil de poèmes, ou les paragraphes réalisant au moins un certain score basé sur les spécificités des unités qu'ils contiennent.

Si le critère de sélection est lié à la présence de certaines unités, alors celles-ci sont mises en évidence typographiquement. Dans certains cas, selon le mode de sélection des extraits, on peut aussi avoir des doublons : un paramétrage permet à l'utilisateur de choisir soit de présenter chaque contexte une seule fois, en signalant les différents cas qu'il réalise, soit d'afficher le contexte pour chaque réalisation du critère.

La métafonctionnalité *Extrait* sert habituellement à travailler sur une représentation réduite du corpus, une lecture sélective du texte. Elle s'articule donc fortement (et généralement hypertextuellement) avec la métafonctionnalité *Texte* pour faciliter un retour à un contexte non limité aux bornes de l'empan. Ses usages typiques sont : (i) la recherche d'exemples, d'attestations, à l'appui d'un document didactique, scientifique, etc. : la sortie peut alors n'être exploitée que partiellement (on choisit le premier extrait satisfaisant) ; (ii) l'analyse systématique d'un phénomène dans le corpus (comme la polysémie d'un mot), en regroupant les extraits en classes correspondant à différents cas de figure de réalisation du phénomène.

3.1.4. Concordance

Un corpus étant fixé, une concordance est la liste de toutes les occurrences d'un focus, (i) alignées verticalement en colonne (nous dirons "empilées"), (ii) entourées de part et d'autre par leur contexte, (iii) munies d'une indication de localisation, et (iv) triées selon un critère pertinent pour l'analyse. L'intérêt de la présentation en concordance est de créer des effets visuels par les tris et la superposition : elle est spécialement appropriée pour l'observation des récurrences et des contrastes au voisinage immédiat du focus, tout en gardant un accès direct à un contexte élargi, par un lien hypertexte renvoyant à la métafonctionnalité *Texte*, et au contexte global, par la mention de la localisation. Ses propriétés de mise en page très particulières, permettent de la distinguer nettement de la métafonctionnalité *Extraits*.

	Métafonctionnalité *Extraits*	Métafonctionnalité *Concordance*
Position du focus	au fil du texte (selon le type d'empan choisi)	centré, aligné verticalement sur une colonne
Disposition du contexte	comme un paragraphe, sans interruption particulière ni alignement, "naturel", comme dans le texte	sur une seule ligne (quitte à équiper la fenêtre d'un ascenseur horizontal), pour ne pas rompre le regroupement vertical des occurrences du focus.
Usage	Travail sur des passages comme unités d'étude ; lecture s'apparentant à une lecture continue, s'appuyant sur une mise en forme usuelle (pas de lignes artificiellement longues comme dans la concordance).	Voisinage immédiat, syntagmatique, orienté, sensibilité à la distance au focus ; mise en évidence de constructions, de leur récurrence et de leurs divergences et variantes ; lecture centrée sur le focus.

3.2. Première famille de fonctionnalités de synthèse : les relevés

3.2.1. Vocabulaire

La métafonctionnalité *Vocabulaire* procède à l'inventaire des unités[8], avec indication de leur fréquence. Cette liste peut être exhaustive ou focalisée (filtre). Un tri alphabétique facilite la recherche d'une unité donnée, et induit certains regroupements morphologiques. Corrélativement, il peut mettre en évidence des lacunes significatives. Le tri hiérarchique (sur la fréquence décroissante) permet de situer les unités dans une gamme de fréquences, des unités dominantes aux hapax. Ses deux atouts sont la simplicité et la réduction opérée. Reposant sur des procédures familières (tri et dédoublonnage), l'interprétation des résultats ne nécessite pas de comprendre un calcul complexe. Et elle fournit une vue réduite du corpus ou d'un phénomène, puisque l'on voit non pas les occurrences (en contexte), mais les types.

La métafonctionnalité *Vocabulaire* joue souvent un rôle de point d'entrée dans l'analyse. Tout d'abord, elle permet une prise de connaissance synthétique du corpus par le balayage systématique de toutes les attestations, et donne des repères pour ajuster un seuil. Elle guide la recherche d'un focus significativement présent dans le corpus et avec un ordre de grandeur de fréquence exploitable, et facilite sa formulation en piochant dans les formes attestées.

3.2.2. Mesures

La métafonctionnalité *Mesures* permet d'obtenir des caractérisations quantitatives, focalisées ou non, à base de décomptes simples (comme effectif, proportion, moyenne). Certaines mesures peuvent être prédéfinies, concernant des tailles (nombre de types, nombre d'occurrences, nombre de hapax,…), la vérification de lois statistiques (table et diagramme de

[8] Pour mémoire, les unités sont considérées sous l'angle d'une propriété donnée : donc on peut lister par exemple les différentes graphies attestées dans le corpus, ou les lemmes, ou les catégories grammaticales, etc.

zipf/pareto), des évaluations globales notamment dans une perspective stylométrique (richesse du vocabulaire, lisibilité).

3.3. Deuxième famille de fonctionnalités de synthèse : les études de positions

Un tableau résume la complémentarité des trois métafonctionnalités concernant les positions :

	Métafonctionnalité *Déroulement*	Métafonctionnalité *Distribution*	Métafonctionnalité *Evolution*
Le corpus est vu comme ayant une structure…	Continue, éventuellement pseudo continue (discrétisation en tranches) Linéaire orientée	Partitionnée Sans orientation ni contiguïtés déclarées	Partitionnée Linéaire orientée
Usage typique	Analyse intratextuelle, approches topologique et topographique	Analyse intertextuelle, contrastive	Analyse chronologique, avec périodes objectivées
Sémiologie graphique	courbe	(histogramme, carte)	histogramme

3.3.1. Déroulement linéaire

Le corpus est ici considéré comme une structure continue -linéaire, syntagmatique-. La métafonctionnalité *Déroulement* vise alors à rendre compte de la régularité ou de l'irrégularité (apparitions groupées, "en rafales") des positions d'un focus (vue focalisée) ou de l'ensemble des unités (vue panoramique). Elle peut également étudier l'évolution continue d'une caractéristique au fil des unités (par exemple, l'accroissement du vocabulaire).

La modélisation peut être pseudo continue : elle mobilise alors une discrétisation, un découpage en « tranches » sans identité propre, au sens où, dans le cadre de cette métafonctionnalité, on ne cherche pas à caractériser ces tranches elles-mêmes.

3.3.2. Distribution

La métafonctionnalité *Distribution* sert à mettre en évidence les affinités (ou les évitements) entre des unités et des parties du corpus. Pour chaque partie, elle indique les unités les plus saillantes, et éventuellement celles qui sont sous-représentées, à l'aune du corpus (plus exactement du fond choisi). Pour le corpus, elle repère les formes de base, à savoir celles qui sont banales dans toutes les parties. Et pour un focus, elle évalue le caractère normal ou remarquable de la fréquence du focus dans les différentes parties du corpus.

La métafonctionnalité *Distribution* s'appuie sur une mesure de la distribution des fréquences dans un corpus partitionné. Certaines mesures sont plus intuitives : fréquence, fréquence relative. D'autres visent une plus grande fiabilité en reposant sur une modélisation statistique : écart-réduit, et surtout spécificités (Lebart & Salem 1994).

3.3.3. Evolution

Par opposition à la métafonctionnalité *Déroulement*, on considère que le corpus se représente comme une succession de périodes individualisables, *a priori* dotées d'une consistance propre. La métafonctionnalité *Evolution* offre alors différents calculs mettant en évidence diverses associations entre la répartition des unités et la structure chronologique (ou équivalente), tels que celui du profil d'un focus (présence croissante, ou décroissante, ou période ou suite de périodes de présence caractéristique), la recherche des formes dont l'accroissement ou l'effacement est le plus significatif au fil du corpus, le repérage des accroissements ou des chutes significatives de l'emploi de certaines unités d'une période à l'autre.

3.4. Troisième famille de fonctionnalités de synthèse : les associations

3.4.1. Séquences

La métafonctionnalité *Séquences*, typiquement implémentée par les segments répétés (Lebart & Salem 1994), vise à repérer des enchaînements récurrents, des figements. Elle opère une synthèse des successions syntagmatiques d'unités en corpus, sans nécessairement préjuger de leur structure. Son usage caractéristique est la reconstitution *a posteriori*, pour l'interprétation voire pour les calculs, d'unités linguistiques non décrites dans la segmentation initiale du corpus[9], et de motifs réguliers de portée plus ou moins longue.

3.4.2. Cooccurrences

La cooccurrence est la présence d'une unité dans le voisinage syntagmatique d'une autre (les voisinages typiques étant de l'ordre de la phrase ou du paragraphe[10]). La métafonctionnalité *Cooccurrences* vise à déceler de telles attirances contextuelles remarquables, au vu du comportement global des unités dans le corpus. Elle peut être orientée (en distinguant, pour deux unités, les deux cas de figures, selon l'unité qui précède l'autre). Dans ses versions statistiques, elle peut être basée sur un calcul de spécificités, ou mobiliser un calcul dédié (Lafon 1981). Elle est utile pour repérer des associations plus souples que celles des *Séquences*.

3.4.3. Analogies

Les *Séquences* et *Cooccurrences* décrivent des associations syntagmatiques. La métafonctionnalité *Analogies* s'intéresse aux similarités, entre unités ou entre parties. Elle peut capter des associations paradigmatiques, en mettant en relation des unités qui voisinent dans le corpus avec les mêmes cooccurrents sans pour autant nécessairement se trouver ensemble dans les mêmes contextes. La métafonctionnalité *Analogies* se base généralement sur un tableau de caractérisation d'unités ou de parties dans un espace de description. Elle peut prendre différentes formes : production d'un tableau de similarité ou de distances, classification non supervisée, visualisation par analyse factorielle ou arborée, par carte de Kohonen. Elle peut fournir des indications tant quantitatives (mesure de cohésion, de distance…) que qualitatives (facteurs concourant au rapprochement ou à la différentiation).

3.4.4. Corrélations

La métafonctionnalité *Corrélations* étudie quant à elle la force et l'orientation du lien éventuel entre deux propriétés, caractérisées par leurs valeurs au fil des occurrences. La statistique propose pour cela des techniques classiques (chi-2, régression…) encore peu intégrées dans les logiciels de textométrie[11].

---

[9] Certaines séquences peuvent aussi quelquefois être recherchées et identifiées à l'aide de ressources dictionnairiques (mots composés, locutions) ; le calcul peut néanmoins être mobilisé pour trouver des séquences d'autres natures.

[10] Comme pour les fonctionnalités *Extraits* ou *Concordance*, les voisinages peuvent être défini par fenêtrage, délimiteur, ou structure englobante.

[11] Cela pourrait tenir à la prise en compte relativement récente des corpus étiquetés, dotant les unités de multiples propriétés.

4. Conclusion : Tradition et innovation

En adoptant un point de vue utilisateur, la typologie proposée ici apporte une structuration pédagogique et mnémonique correspondant aux types de questionnements de l'analyse textuelle : caractérisation d'une unité ou d'un texte, affinités entre les unités et les parties du corpus, voisinages des unités, etc. Cette typologie renouvelle et unifie la vision des fonctionnalités textométriques. Par exemple, la métafonctionnalité *Déroulement* groupe ce qui est habituellement dispersé : rafales, cartes des sections, accroissement du vocabulaire, topologie... De plus, cette métafonctionnalité se trouve bien correspondre aux recherches actuelles des textomètres en topologie ou topographie (Salem & Mellet 2008), notre typologie s'ajuste donc naturellement à cette problématique existante.

Par ailleurs, élaborée dans le contexte de la conception d'un nouveau logiciel (projet Textométrie), cette typologie couvre des aspects encore peu développés, comme une intégration forte des corpus étiquetés (définition transverse aux calculs des propriétés d'analyse et d'affichage, construction de propriétés), ou encore le besoin d'un environnement outillé d'analyse et d'aide à l'interprétation des résultats (organisation par filtrages et regroupements, vues synoptiques et instruments de comparaison). Ceci étant, ces propositions innovantes restent clairement au service de la valorisation des calculs au cœur de la textométrie, patrimoine scientifique riche et bien vivant.

Cette communication a été préparée dans le cadre du projet Textométrie ANR-06-CORP-029 ; elle a bénéficié d'une réflexion collective dépassant le cercle des auteurs ayant rédigé ces lignes.

Références

Brunet E. (2006) – *Hyperbase, Logiciel documentaire et statistique pour la création et l'exploitation de bases hypertextuelles, Manuel de référence*, Institut de linguistique française, "Bases, Corpus et Langage", Université de Nice, mai 2006, 151 pages.

Fleury S. (2009) – *Le Métier Textométrique, aka Le Trameur, Manuel d'utilisation*, Centre de Textométrie, CAT2T, Université de Paris 3, juillet 2009, 127 pages.

Fracchiolla B., Kuncova A., Maisondieu A. (2003) – *Lexico 3, outils de statistique textuelle, Manuel d'utilisation*, Version 3.41, SYLED-CLA2T, Université Paris 3, février 2003, 50 pages.

Heiden S. (2002). *Weblex. Manuel Utilisateur*. Version 4.1, Laboratoire ICAR, UMR 5191, ENS Lyon, janvier 2002, 180 pages.

Lafon P. (1981) – Analyse lexicométrique et recherche des cooccurrences, *Mots*, 3, 95-148.

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

Leblanc J.-M. (2005) - *Les voeux présidentiels sous la Cinquième République (1959-2001). Recherches et expérimentations lexicométriques à propos de l'ethos dans un genre discursif rituel*, Thèse de Doctorat, Sciences du langage, Université de Paris 12, 8 décembre 2005.

Salem A. and Fleury S. (éds) (2007) - Explorations textométriques, *Lexicometrica*..

Salem A. and Mellet S. (éds) (2008) – Topographie et topologie textuelles, *Lexicometrica*

Sites internet

Projet Textométrie : http://textometrie.ens-lsh.fr/ *Logiciel Xaira* : http://www.xaira.org/

Portail et revue Lexicometrica : http://www.cavi.univ-paris3.fr/lexicometrica/.

Textopol, Ressources informatisées pour l'analyse du discours politique (Pierre Fiala, Jean-Marc Leblanc) : http://textopol.org/