L'approche lexicométrique dans un commentaire de texte

Jean-Marc Veran – Mai/juin 2017

Introduction

Nous proposons dans les quelques pages qui suivent de présenter une méthodologie d'analyse des textes qui se fonde sur les acquis de l'approche lexicométrique dans le cadre plus général de l'analyse du discours.

Sans rentrer dans les détails, l'analyse du discours est une discipline qui a vu le jour à la suite des travaux de Harris Zellig dans les années 50 du siècle dernier et qui ambitionnait d'aborder le texte en terme de communication. Alors que la linguistique, essentiellement descriptive, faisait du texte un « objet clos », l'analyse du discours visait à le comprendre aussi bien dans son articulation interne qu'au sein du contexte dans lequel il avait été produit.

Initialement, son objet concernait plus particulièrement le discours politique ou social. Depuis, son champ d'étude s'est élargi et aujourd'hui il touche toute forme d'énoncés aussi bien écrit qu'oral. De plus, le développement des moyens informatiques lui a permis de toucher à des corpus de plus en plus importants. Il existe un large éventail d'approches de l'analyse d'un discours (énonciative, communicationnelle, pragmatique, sémiotique…), mais, pour notre part, nous nous attacherons à l'approche qualifiée de lexicométrique.

Cette approche des textes, dans une formulation très certainement réductrice et peut-être un peu brutale, pose que le vocabulaire « trahit » son sens. Ainsi, à partir de l'analyse quantitative des occurrences des différents vocables, de leur position, de leur environnement… le texte se dévoile progressivement. Se fondant sur un postulat l'on pourrait douter de la validité de la méthode, mais les résultats qu'apporte son utilisation depuis une cinquantaine d'années viennent plutôt dire le contraire. Il y a certes un certain nombre de détracteurs de la lexicomètrie qui invalident sa méthode mais la question tient plus probablement à la lemmatisation préalable du texte qu'à l'analyse proprement dite.

Que signifie ce terme de « lemmatisation » ? Il s'agit du regroupement des mots selon leur forme canonique (c'est-à-dire la forme dans laquelle ils apparaissent dans un dictionnaire). Ainsi « homme » est le lemme de « homme » et « hommes », « humain » celui de « humain », « humaine », « humains » et « humaines », enfin « aller » celui de toutes les formes conjugées à toutes les personnes, tous les modes et tous les temps du verbe

« aller » : « vais », « vas », « va », « allons »… La lemmatisation est une opération fastidieuse lorsqu'elle est faite manuellement mais pouvant occasionner des erreurs particulièrement préjudiciables lorsqu'elle est faite automatiquement. Ainsi, il n'est pas nécessairement évident pour un ordinateur de distinguer entre le substantif « parti » (par exemple un parti politique) et « parti », forme conjugée du verbe « partir ». De grands progrès ont cependant été réalisés ces dernières années et les logiciels de lemmatisation, sans atteindre le sans faute, présentent cependant des taux de réussite relativement satisfaisants.

Le texte est ainsi transformé en une base de données constituée de termes sur lesquels vont être appliqués un certain nombre de procédures permettant de mettre en évidence le nombre d'occurrences de chaque vocable, procéder à différentes analyses comparatives ou thématiques…

Il ne s'agit pas dans ces quelques lignes de faire un exposé détaillé de ces méthodes qui nécessitent souvent des compétences mathématiques de bon niveau, voire de très bon niveau, mais plus modestement nous proposons d'approcher cette méthodologie au moyen de techniques qui s'en inspirent, du moins qui visent le même but, mais sans nécessiter des connaissances mathématiques avancées.

De plus et afin d'en faciliter la compréhension nous l'aborderons par une série d'exemples appartenant au champ de l'exégèse biblique, du magistère de l'Église ou d'ouvrages théologiques.

En préalable trois remarques paraissent nécessaires : dans la mesure où elle se fonde sur l'analyse statistique du vocabulaire, elle est d'autant plus efficace que le texte est long. Le résultat sur des textes de mille à deux mille mots apparaît parfois décevant. La raison en est simple, plus le texte est court plus la discrimination statistique des mots entre eux est faible. Mais notons-le, sur un texte court nous saisissons plus aisément les mouvements du texte et les mots clés apparaissent assez facilement.

La deuxième remarque est relative à l'utilisation des moyens informatiques. Si le développement de cette méthode d'analyse des textes doit beaucoup à l'informatique, il ne faudrait cependant pas croire qu'il suffit d'appuyer sur un bouton pour voir jaillir le sens. Son application nécessite un travail souvent fastidieux et en aucun cas ne nous affranchit d'un effort d'analyse des résultats obtenus, ni de la lecture attentive du ou des documents.

La troisième remarque tient à la cohérence du texte sur lequel est appliquée cette méthode d'analyse. L'idéal est de travailler sur un texte qui présente une certaine unité en termes de temps et de lieu d'élaboration mais également d'auteur et de thématique. Cette précision peut paraître insolite et exige quelques précisions. Un bon exemple pour saisir notre propos est sans doute celui du livre des Psaumes. Si ces textes présentent une certaine unité thématique, ils proviennent cependant de différents auteurs et leur composition s'étale sur une période relativement longue. Il s'en suit que s'il existe bien un corpus de vocables qui traverse l'ensemble des unités textuelles qui les composent,

l'analyse devra vérifier leur cohérence en terme de signification selon les auteurs et selon les périodes au risque de créer un biais dans l'interprétation. Notons que ce travail n'est pas toujours facile lorsque l'on est au fait des problèmes de datation des Psaumes.

Il est donc toujours essentiel de mener une réflexion sur le texte avant de s'engager dans ce type de travail.

1er exemple : L'Epitre aux Romains

Ce premier exemple, volontairement simple, permet déjà de situer la problématique à laquelle nous nous attachons. Le texte de Romains est bien connu et les difficultés qu'il soulève également. Il ne s'agit pas ici d'en faire un commentaire mais de montrer qu'une simple analyse du vocabulaire donne la possibilité d'en situer les enjeux essentiels et de montrer les points incontournables auxquels toute exégèse doit au moins s'attacher. Au niveau auquel nous nous situerons nous ne ferons pas de découvertes particulières, d'ailleurs notre objet n'est pas là. Notre visée est simplement de montrer qu'une analyse lexicomètrique, même sommaire, ne laisse guère échapper les sujets abordés ou les thèmes dominants.

Nous l'avons dit plus haut, toute analyse lexicomètrique débute par la lemmatisation du texte. Par chance, tous les textes bibliques, qu'ils appartiennent au corpus vétéro ou néo-testamentaires, ont été lemmatisés. On trouvera l'ensemble du Nouveau Testament lemmatisé à l'adresse suivante :

https://github.com/morphgnt/tischendorf/downloads

Il s'agit de la version de Tischendorf car la version Nestlé-Aland (NA27) est soumise à un copyright et ne peut donc être utilisée sans autorisation. Elle n'en diffère cependant que de peu car ces deux bases utilisent des textes communs, ceux des grands onciaux des IVème et Vème siècles (Sinaïticus, Vaticanus, Alexandrinus…). A l'état brut, elle est difficilement utilisable et nécessite d'être retravaillée pour être d'une utilisation facile. Sa transformation n'est pas insurmontable pour celui qui possède quelques connaissances informatiques en matière de tableur. Le vocable est mentionné juste après la référence, à sa suite l'on trouve son analyse grammaticale et son numéro Strong puis le lemme, qui est indiqué à la fin de la ligne. En référant au texte grec, le codage des caractères ne soulève pas de difficultés particulières.

Voici les 2 premiers versets de Romains dans la base « morphgnt » :

RO 1:1.1 C *PAU=LOS *PAU=LOS N-NSM 3972 *PAU=LOS ! *PAU=LOS

RO 1:1.2 . DOU=LOS DOU=LOS N-NSM 1401 DOU=LOS ! DOU=LOS [1II]1

RO 1:1.3 . *XRISTOU= *XRISTOU= N-GSM 5547 *XRISTO/S ! *XRISTO/S

RO 1:1.4 . *)IHSOU=, *)IHSOU=, N-GSM 2424 *)IHSOU=S ! *)IHSOU=S

RO 1:1.5 . KLHTO\S KLHTO\S A-NSM 2822 KLHTO/S ! KLHTO/S

RO 1:1.6 . A)PO/STOLOS A)PO/STOLOS N-NSM 652 A)PO/STOLOS ! A)PO/STOLOS

RO 1:1.7 . A)FWRISME/NOS A)FWRISME/NOS V-RPP-NSM 873 A)FORI/ZW ! A)FORI/ZW

RO 1:1.8 . EI)S EI)S PREP 1519 EI)S ! EI)S

RO 1:1.9 . EU)AGGE/LION EU)AGGE/LION N-ASN 2098 EU)AGGE/LION ! EU)AGGE/LION

RO 1:1.10 . QEOU=, QEOU=, N-GSM 2316 QEO/S ! QEO/S

RO 1:2.1 . O(\ O(\ R-ASN 3739 O(/S ! O(/S

RO 1:2.2 . PROEPHGGEI/LATO PROEPHGGEI/LATO V-ADI-3S 4279 PROEPAGGE/LLOMAI ! PROEPAGGE/LLW

RO 1:2.3 . DIA\ DIA\ PREP 1223 DIA/ ! DIA/

RO 1:2.4 . TW=N TW=N T-GPM 3588 O(! O(

RO 1:2.5 . PROFHTW=N PROFHTW=N N-GPM 4396 PROFH/THS ! PROFH/THS

RO 1:2.6 . AU)TOU= AU)TOU= P-GSM 846 AU)TO/S ! AU)TO/S

RO 1:2.7 . E)N E)N PREP 1722 E)N ! E)N

RO 1:2.8 . GRAFAI=S GRAFAI=S N-DPF 1124 GRAFH/ ! GRAFH/

RO 1:2.9 . A(GI/AIS, A(GI/AIS, A-DPF 40 A(/GIOS ! A(/GIOS

Et après traitement de ceux-ci l'on obtient :

	Livre	Ch.	Ver.	Place	Vocable	Lemme	Analyse grammaticale
1	Rm	1	1	1	**Pau=loj**	**Pau=loj**	Nom Nominatif Masculin Singulier
2	Rm	1	1	2	**dou=loj**	**dou=loj**	Nom Nominatif Masculin Singulier
3	Rm	1	1	3	**Xristou=**	**Xristo/j**	Nom Génitif Masculin Singulier
4	Rm	1	1	4	**)Ihsou=**	**)Ihsou=j**	Nom Génitif Masculin Singulier
5	Rm	1	1	5	**klhto\j**	**klhto/j**	Adjectif Nominatif Masculin Singulier
6	Rm	1	1	6	**a)po/stoloj**	**a)po/stoloj**	Nom Nominatif Masculin Singulier
7	Rm	1	1	7	**a)fwrisme/noj**	**a)fori/zw**	Verbe Parfait Participe Passif Nominatif Masculin Singulier
8	Rm	1	1	8	**ei)j**	**ei)j**	Préposition
9	Rm	1	1	9	**eu)agge/lion**	**eu)agge/lion**	Nom Accusatif Neutre Singulier
10	Rm	1	1	10	**qeou=**	**qeo/j**	Nom Génitif Masculin Singulier
11	Rm	1	2	1	**o(**	**o(/j**	Pronom relatif Accusatif Neutre Singulie
12	Rm	1	2	2	**proephggei/lato**	**proepagge/llw**	Verbe Aoriste Indicatif Moyen déponent 3° personne du singulier
13	Rm	1	2	3	**dia**	**dia/**	Préposition
14	Rm	1	2	4	**tw=n**	**o(**	Article Génitif Masculin Pluriel
15	Rm	1	2	5	**profhtw=n**	**profh/thj**	Nom Génitif Masculin Pluriel
16	Rm	1	2	6	**au)tou=**	**au)to/j**	Pronom personnel Génitif Masculin Singulier
17	Rm	1	2	7	**e)n**	**e)n**	Préposition
18	Rm	1	2	8	**grafai=j**	**grafh/**	Nom Datif Féminin Pluriel
19	Rm	1	2	9	**a(gi/aij**	**a(/gioj**	Adjectif Datif Féminin Pluriel

Il est dès lors très simple en réalisant un tri sur la colonne « Lemme » de déterminer le nombre d'occurrences de chacun. Voici le résultat après classement par nombre d'occurrences des principaux substantifs. Nous les avons rangés en deux colonnes tant cette présentation est suggestive :

Classement	Vocables	Nb occ.
1	Θεός	153
3	χριστός	65
5	Κύριος	43
7	Ἰησους	36
9	Πνεύμα	34

Classement	Vocables	Nb occ.
2	νομός	74
4	ἁμαρτία	48
6	πίστις	40
8	δικαιοσύνη	34
10	ἔθνος	29

(Θεός = Dieu, χριστός = Christ, Κύριος = Seigneur, Ἰησους = Jésus, Πνεύμα = Esprit ou esprit, νομός = loi, ἁμαρτία = péché, πίστις = foi, δικαιοσύνη = justice, ἔθνος = nation)

Ce simple classement permet de dégager les grands thèmes de la lettre. Dieu et l'Esprit encadrent Jésus et ses titulatures de Christ et Seigneur tandis que les cinq vocables loi, péché, foi, justice et nation nous livrent les principales problématiques théologiques de la lettre.

Nous n'irons pas plus loin dans cet exemple, il s'agissait uniquement de montrer qu'un simple classement des principaux substantifs d'un texte permet de déterminer, sans grand risque et sans passer à côté d'un thème important, les problématiques qui le commandent.

Pour information voici les dix suivants qui viennent largement compléter les problématiques exposées par Paul aux Romains :

a;nqrwpoj	27	homme
sa,rx	26	chair
ca,rij	24	grâce
qa,natoj	22	mort
avdelfo,j	19	frère
do,xa	16	gloire
e;rgon	15	œuvre
kardi,a	15	cœur
peritomh,	15	circoncision
path,r	14	père

2ème exemple : Les mots et leur environnement

Dans notre démarche précédente nous ne nous étions guère souciés de l'environnement dans lequel chacun des principaux vocables se trouvaient. Il s'agit ici d'analyser le contexte dans lequel ils sont employés, c'est-à-dire les mots avec lesquels ils sont liés dans une phrase. Là encore, nous nous limiterons aux substantifs mais l'on pourrait étudier l'ensemble des verbes qui leur sont associés ou toute autre question similaire. Nous prendrons comme exemple le Livre du Siracide. C'est un texte dont la taille importante - il ne contient pas moins de 51 chapitres et 18 424 mots dans sa version grecque - permet des statistiques lexicales assez poussées. Nous avons retenu le texte établi par Alfred Rahlfs dans son édition de la LXX (Septante) que l'on trouvera sur :

http://ccat.sas.upenn.edu/gopher/text/religion/biblical/lxxmorph/36.Sirach.mlxx

Là encore, la base de données demande d'être retraitée pour faciliter son utilisation.

Dans un premier temps, et comme précédemment, on classe les substantifs par nombre d'occurrences dans le texte. Voici les dix premiers.

ku/rioj	201	Seigneur
a)/nqrwpoj	118	homme/humain
kardi/a	86	cœur
yuxh/	80	souffle vital
a)nh/r	79	homme/mari
lo/goj	70	parole
h(me/ra	66	jour
e)/rgon	65	œuvre
kairo/j	63	temps fixé/moment
xei/r	59	main

Pour notre exemple nous ne retiendrons que les cinq premiers. Il serait possible d'en sélectionner plus mais, en la matière, un trop grand nombre de mots risque, au moins au début d'une analyse, de se révéler difficile à gérer. Rien n'interdit dans un second temps d'aller plus loin. Ainsi, comme nous le verrons, dans notre cas les cinq premiers vocables livrent déjà une information qui est loin d'être négligeable.

La deuxième opération consiste alors, pour chacun de ces cinq vocables, à faire l'inventaire de tous les substantifs qui leur sont associés dans chacun des versets dans lesquels ils apparaissent. Ainsi le vocable de « Seigneur » se rencontre dans 189 versets et dans 21 de ces versets on trouve le mot « crainte », dans 17 le mot « sagesse »… Ce type de recherche n'est pas très difficile à mettre en œuvre sous Excel.

κύριος	189	ἄνθρωπος	109	καρδία	84	ψυχή	78	ἀνήρ	76
φόβος	21	κύριος	14	κύριος	15	κύριος	9	γυνή	11
σοφία	17	ἀνήρ	11	ἄνθρωπος	8	καρδία	7	ἄνθρωπος	11
ἄνθρωπος	17	καρδία	8	ψυχή	7	ἀνήρ	7	ψυχή	7
δόξα	16	ἔργον	8	λύπη	7	ζωή	6	κύριος	5
καρδία	13	ζωή	7	φόβος	7	πρόσωπον	5	καρδία	5

Seigneur	189	humain	109	cœur	84	Souffle vital	78	mari	76
crainte	21	Seigneur	14	Seigneur	15	Seigneur	9	femme	11
sagesse	17	mari	11	humain	8	cœur	7	humain	11
humain	17	cœur	8	Souffle vital	7	mari	7	souffle vital	7
gloire	16	œuvre	8	tristesse	7	vie	6	Seigneur	5
cœur	13	vie	7	crainte	7	visage	5	cœur	5

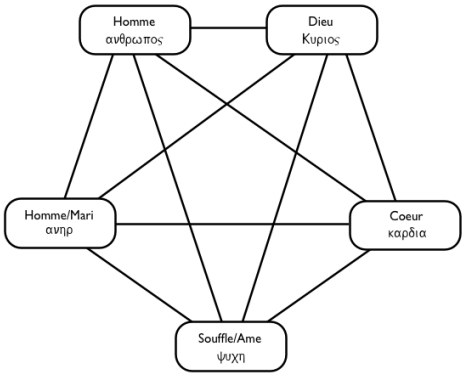
(Les différences entre nombre d'occurrences et nombre de versets dans lesquels l'occurrence apparaît provient de sa répétition)

L'analyse se fait en deux temps, d'abord les relations au sein du groupe des cinq premiers vocables puis les relations avec les autres termes.

ku/rioj	189	**a)/nqrwpoj**	118	**kardi/a**	86	**yuxh/**	80	**a)nh/r**	79
		κύριος	14	κύριος	15	κύριος	9		
		ἀνήρ	11	ἄνθρωπος	8	καρδία	7	ἄνθρωπος	11
ἄνθρωπος	17	καρδία	8	ψυχή	7	ἀνήρ	7	ψυχή	7
								κύριος	5
καρδία	13							καρδία	5

Comme le montre le tableau ci-dessus, les cinq premiers vocables sont fortement intercorrélés, chacun entretenant un lien avec chacun. Ainsi se dessine un système de relations que nous pouvons représenter à partir d'un schéma pentagonal. On objectera que ce schéma n'est que la conséquence de la prise en compte des 5 premiers substantifs et que si nous avions sélectionné les six premiers nous aurions obtenu un schéma hexagonal. En fait cela n'est pas le cas. Lorsque nous augmentons le nombre de substantifs, un nouvel ensemble se dessine qui, bien que relié au premier au travers des substantifs κύριος (Seigneur) et ἄνθρωπος (homme), se déploie de manière indépendante de celui-ci. Nous l'évoquerons un peu plus loin.

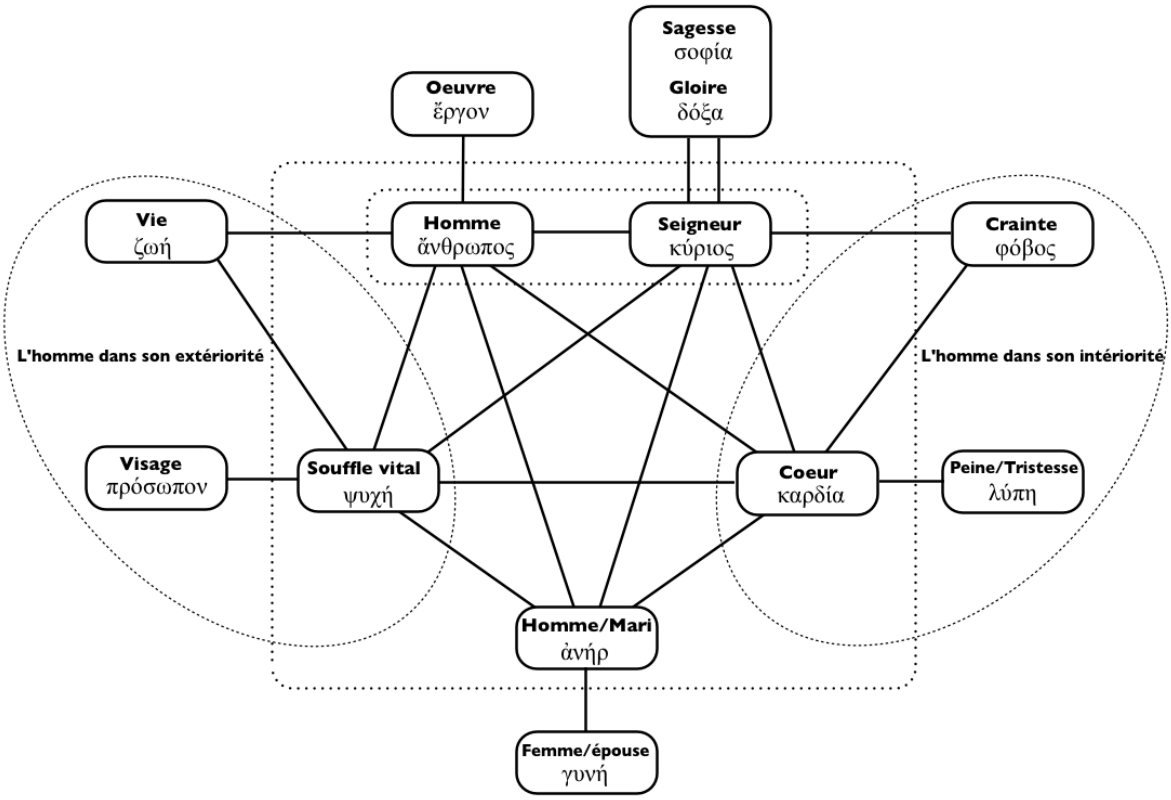
Le schéma de base est donc le suivant :



Le second tableau (ci-dessous) permet de distinguer les relations les plus significatives avec les autres occurrences de vocabulaire n'appartenant pas au premier groupe.

ku/rioj 189		**a)/nqrwpoj** 118		**kardi/a** 86		**yuxh/** 80		**a)nh/r** 79	
φόβος	21							γυνή	11
σοφία	17								
δόξα	16	ἔργον	8	λύπη	7	ζωή	6		
		ζωή	7	φόβος	7	πρόσωπον	5		

Il suffit alors de rajouter ces nouveaux liens au schéma précédent et l'on obtient le schéma suivant :



Nous pourrions presque dire que ce schéma parle de lui-même. En effet, il permet d'appréhender les principaux éléments qui président à l'articulation de la pensée de l'auteur et offre ainsi une première vision synthétique de l'œuvre de Ben Sira. Dieu et l'homme en sont les pivots. Dieu caractérisé par sa sagesse et sa gloire et l'homme compris dans ses dimensions intérieures (cœur, crainte, tristesse), vitales (souffle vital, vie, visage) et familiales (mari, épouse) et qui, dans chacune de ses composantes, est lié à Celui qui l'a fait, formant ainsi comme un tout avec Lui.

Le commentaire qui suivra d'un tel schéma sera des plus classiques. Il débutera par une analyse de chacun des vocables en vérifiant pour chacun si son sens correspond chez Ben Sira à celui que l'on rencontre généralement dans le corpus biblique et en référant, selon les cas, à un ou plusieurs ouvrages d'anthropologie biblique ou traitant spécifiquement du sujet. Puis il s'agira de collationner les versets correspondant à chacune de ces relations, faire les regroupements auxquels nous invitent les liens qui ont été établis et les commenter sur la base de l'analyse menée au stade précédent. Enfin, une synthèse dont l'articulation se fera sur la base du schéma initial.

Comme nous l'avons évoqué plus haut, il est possible d'aller plus loin dans cette analyse en élargissant le champ des vocables pris en compte. Nous pourrions alors voir apparaître le rapport de l'homme à ses œuvres et au temps avec les vocables de « parole », « jour », « œuvre », « instant » et « main ». Ainsi sera mise en relief une double perspective qui commande toute l'œuvre de Ben Sira. De part et d'autre du Seigneur qui occupe la place centrale, d'un côté l'homme dans son être, que nous venons d'évoquer et, de l'autre,

l'homme dans son faire (mains) et dans son dire (parole) au travers de ses œuvres (œuvre) dans le temps (jour et instant).

3ème exemple : Le livre des Proverbes

Avec ce troisième exemple nous abordons le texte hébreu. Ce type d'analyse est évidemment réservé à ceux qui possèdent un logiciel biblique car à notre connaissance il n'existe pas de base de données lémmatisées du texte massorétique libre de droit sur Internet.

Si nous avons choisi cet exemple, c'est afin de mettre en évidence les précautions qu'il est souvent nécessaire de prendre avant de s'engager dans ce type d'analyse.

Le résultat brut du dénombrement des nombres d'occurrences du texte des Proverbes donne les résultats suivants pour les cinq premières :

Ble	97	cœur
vya	91	homme
hwhy	86	Seigneur
%r,D,	80	chemin
!Be	79	fils

De la même manière que précédemment, l'on pourra établir le tableau présentant l'inventaire des mots associés à chacune de ces cinq premières occurrences :

cœur		homme		Seigneur		chemin		fils	
Seigneur	12	chemin	17	crainte	14	homme	14	père	14
humain	11	cœur	10	chemin	13	Seigneur	13	cœur	8
homme	11	lèvres	9	horreur	13	cœur	9	mère	7
lèvres	9	Seigneur	9	cœur	12	voie	8	parole	7
chemin	8	œil	7	œil	9	œil	5	discipline	6

Ce tableau permet de déjà bien situer l'enjeu du texte des Proverbes. Il soulève toutefois quelques interrogations.

Ainsi le mot « cœur » qui présente en hébreu deux orthographes : Ble mais également bb'le. Il y aurait donc lieu de rajouter les occurrences de bb'le à celle de Ble. En outre, comme nous pouvons le voir, nous sommes en présence de deux vocables à propos de l'homme : vya traduit par « homme » et ~d'a' par « humain ». Doit-on les regrouper ou bien les distinguer, d'autant que « humain » étant présent quarante cinq fois dans le texte, cela n'est pas neutre en terme de classement ? Le même type de remarque peut avoir lieu

pour %r,D, « chemin » et xr;ao « voie ». Enfin quelle attitude adopter envers le vocable ~yhil{a/ (Dieu) vis-à-vis du nom « Seigneur » ? Notons cependant qu'il n'apparaît que cinq fois et ne modifie guère l'ordre des occurrences comme cela est le cas avec « homme » et « humain ».

Cet exemple avait uniquement pour but de bien montrer qu'il est toujours nécessaire de préciser les conditions dans lesquelles se font les décomptes au risque d'abuser son lecteur et s'abuser soi-même.

4ème exemple : Comparaison de trois textes du magistère pontifical

Jusqu'à maintenant nous avons appliqué notre méthodologie d'analyse à des textes pour lesquels nous disposions de la base de données lemmatisées. Avec ce troisième exemple, nous allons aborder non seulement une nouvelle technique qui vise à comparer trois textes mais également des textes pour lesquels nous ne disposons pas de base de données lemmatisées. Il s'agira de confronter le texte français de trois documents du Magistère pontifical à l'occasion de l'anniversaire décennal de Rerum Novarum de Léon XIII afin de mettre en lumière les inflexions et les permanences du discours magistériel concernant la doctrine sociale de l'Eglise.

Notons que ce type d'analyse est très proche de ce qui fut mis en œuvre aux origines de la lexicomètrie. En effet, c'est au moyen de méthodes assez similaires que certains chercheurs français s'attachèrent à analyser les mutations du discours dans le champ socio-politique. Ils comparaient le vocabulaire employé et son évolution parmi les hommes politiques et les leaders syndicaux de l'époque.

Les trois textes pris en compte sont : Mater et Magistra de Jean XXIII, Octogesima adveniens de Paul VI et Centesimus annus de Jean-Paul II.

Si nous posons le postulat que le vocabulaire « trahit » le sens, alors il s'agit de comparer entre-eux le vocabulaire de chacun des pontifes afin d'y déceler ce qu'ils ont en commun et ce qui les différentie.

Premier temps : constitution de la base de données

Le premier travail, et le plus fastidieux, est d'établir la liste des vocables employés et la fréquence de leur apparition dans le texte. La méthode la plus contraignante mais sans nul doute la plus sûre est de réaliser cette opération soi-même. Pour cela il faut transformer le texte en une suite de termes indépendants puis de les disposer sur une colonne sous Excel. On procède alors à un tri alphabétique et l'on réalise le décompte de chaque occurrence. Enfin on procède aux regroupements lemmatiques qui s'imposent. C'est un travail relativement long mais pas très compliqué.

On peut aussi utiliser les moyens que fournit internet. Ainsi sur HTTP://www.IntraText.com/Catalogo/ (faire la recherche par auteur) on trouvera

quelques textes lemmatisés. Mater et Magistra est situé dans la rubrique consacrée à Jean XXIII et Centesimmus annus dans celle de Jean-Paul II, mais malheureusement pas Octogesima Adveniens. Il faudra cependant être prudent car ce n'est pas un décompte par lemme. En effet, IntraText considère comme différent le même mot s'il est au singulier ou au pluriel (idem pour les verbes, à chaque temps, mode …). Cette base demande donc un travail de regroupement assez fastidieux. Il existe également le logiciel « Tropes » qui procède à la lemmatisation des textes et fournit les nombres d'occurrences de chaque adjectif, substantif et verbe. Les décomptes sont relativement corrects. Ainsi l'adjectif « économique » qu'il repère 129 fois est présent 134 fois tout comme « social » dont il donne 105 occurrences contre 120. En revanche, il détecte bien les 17 occurrences de « économie ».

Cela étant, nous touchons là à un problème assez similaire à celui que nous avons rencontré au paragraphe précédent. Dans Mater Magistra, le substantif « économie » est présent 17 fois, en revanche « économique » ou « économiques » sont présents 134 fois et l'adverbe « économiquement » (non décompté dans Tropes) 14 fois. Dans ces conditions, doit-on ignorer les adjectifs et les adverbes ? Probablement pas. Cependant, il faudra faire preuve de prudence car si le substantif « vérité » et l'adjectif « vrai » appartiennent au même champ sémantique et méritent d'être regroupés, l'adverbe « vraiment » ne le nécessite pas compte tenu qu'il ne renvoie pas nécessairement à une phrase en lien avec une problématique liée à la vérité mais s'emploie le plus souvent pour renforcer une affirmation. Il est donc essentiel, avant de décider de la gestion de la base lemmatisée, de vérifier que l'on n'introduit pas un biais important dans l'analyse. Cette réflexion préalable est essentielle et quelques vérifications sont indispensables avant de s'engager dans ce type d'analyse. C'est sans doute sur cet aspect de la méthode que « l'intelligence humaine » surpasse « l'intelligence informatique » !

On trouvera ci-dessous le tableau des 30 premières occurrences (substantif, adjectif, adverbe et verbe[1]) de chacun des textes que nous avons évoqués. La première colonne indique le classement par nombre d'occurrences, la seconde leur nombre et la troisième le vocable.

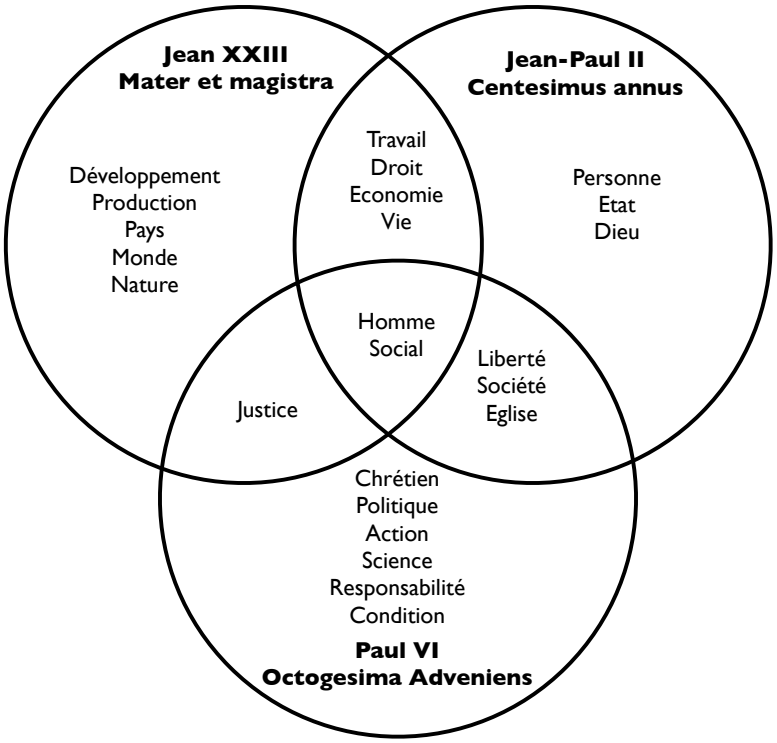
Mater et magistra			Octogesima Adveniens			Centesimus annus		
1	165	économie	1	139	homme	1	313	homme
2	143	social	2	60	chrétien	2	136	social
3	142	hommes	3	55	social	3	126	travail
4	100	vie	4	42	liberté	4	105	personne
5	98	développement	5	41	société	5	103	droit
6	78	travail	6	35	politique	6	98	économie
7	69	production	7	32	action	7	89	vie
8	65	droit	8	30	justice	8	81	Eglise
9	59	pays	9	29	science	9	79	liberté
10	57	justice	10	28	responsabilité	10	73	société
11	54	monde	11	26	condition	11	71	Etat
12	47	nature	12	25	Eglise	12	67	Dieu

[1] A titre d'exemple : le lemme « travail » regroupe les substantifs « travail », « travaux», « travailleur » et « travailleurs » (travailleuse n'est pas présent) et le verbe « travailler ».

13	44	biens	13	25	développement	13	62	justice
14	42	principes	14	22	économie	14	57	développement
15	41	public	15	21	service	15	56	monde
16	41	peuple	16	21	droit	16	54	vérité
17	41	Eglise	17	19	vie	17	51	pays
18	40	entreprise	18	19	idéologie	18	49	besoins
19	40	personne	19	19	monde	19	49	politique
20	40	esprit	20	19	aspiration	20	46	condition
21	39	secteur	21	18	situation	21	44	dignité
22	38	temps	22	17	problème	22	44	nature
23	37	Dieu	23	16	Dieu	23	42	encyclique
24	36	politique	24	16	progrès	24	42	production
25	35	ordre	25	15	conscience	25	39	culture
26	35	ordre	26	15	esprit	26	38	propriété
27	34	action	27	14	nature	27	38	guerre
28	34	moyens	28	13	foi	28	37	temps
29	34	technique	29	10	cœur	29	36	biens
30	33	progrès	30	10	domaine	30	35	conception

Second temps : l'analyse

Si nous prenons les 12 premiers mots de chacun de ces documents, c'est-à-dire les thèmes qui prédominent dans les textes, on peut les classer en 7 catégories. Les « lemmes » communs aux trois pontifes (une catégorie), ceux communs à deux pontifes pris deux à deux (trois catégories) et enfin ceux présents chez un seul pontife (trois catégories). L'on peut alors représenter cela sous forme d'un schéma et l'on obtient la figure suivante :



Ce schéma est assez évocateur des problématiques communes et particulières des ces trois documents. Dans la pratique il suffira alors d'en faire un commentaire en référant aux diverses phrases dans lesquelles ces lemmes se rencontrent. Voilà quelques lignes qui sont très générales et loin d'épuiser tout ce que l'on pourrait dire.

Au cœur de la problématique de ces trois documents et qui est commun aux trois pontifes, il y a l'homme dans sa dimension sociale (on pouvait d'ailleurs si attendre compte tenu du sujet). Chez Jean XXIII la prise de conscience de la dimension mondiale des problématiques se met en place dans le discours pontifical au travers des mots de « monde » et « pays » tandis que la situation de croissance qui caractérise cette époque est bien mise en valeur avec les termes de « production » et « développement ». Chez son successeur, Paul VI, la situation devient très différente. C'est la responsabilité du chrétien dans sa dimension politique et agissante (action) qui se détache avec une particulière netteté et que l'on ne peut pas ne pas mettre en rapport avec les problématiques de l'époque, d'une part, l'émergence de l'individualisme et, d'autre part, la mise en cause du modèle occidental de croissance mais, sans oublier l'influence du travail conciliaire. Enfin avec Jean-Paul II, on assiste à deux phénomènes. D'abord un retour à une partie du vocabulaire de Jean XXIII. Alors qu'avec Paul VI le politique avait pris le pas sur l'économique, celui-ci revient sur le devant de la scène. Mais nous assistons à un second phénomène tout à fait caractéristique, à savoir l'émergence d'un ensemble de termes proches du langage théologique : « personne », « Dieu » associés au vocable « Église ». Ce dernier vocable est tout à fait intéressant à observer. Chez Jean XXIII il occupe la 17ème place, avec Paul VI il passe à la 12ème et avec Jean-Paul II à la 8ème, c'est presque un programme... Même si le vocable de personne, chez Jean-Paul II, désigne l'homme, il est clair que celui-ci ne s'envisage plus seulement dans sa dimension sociale de membre de la communauté internationale mais bien dans une perspective qui relève plus du théologique que du sociologique. Ainsi se développe un mouvement ascendant tout à fait remarquable. A l'homme d'abord perçu comme être social appartenant au monde avec Jean XXIII, succède, avec Paul VI, un homme compris comme chrétien présent au monde et assumant ainsi une responsabilité vis-à-vis de la société qui l'entoure pour, ensuite, dans ce mouvement ascendant, être envisagé avec Jean-Paul II au travers du concept de personne dans une perspective proprement théologique. Les vocables conceptuels ne sont pas en reste. « Droit » et « justice » chez Jean XXIII, « justice » et « liberté » chez Paul VI enfin « liberté » et « droit » chez Jean-Paul II. Il y a là tout un champ de réflexion qui s'ouvre. Et si nous avions poussé l'analyse un peu plus loin, nous aurions pu voir comment le vocable de « vérité » très peu présent chez Jean XXIII, progresse chez Paul VI, pour finalement occuper la 16ème place chez Jean-Paul II et devenir un vocable central dans le discours de Benoît XVI.

À ce stade d'analyse, somme toute assez sommaire, beaucoup de choses sont déjà dites. Tout ne l'est pas, nous l'accordons, mais l'essentiel est là et il n'est pas certain que ce qui nous est apparu au travers de cette démarche aurait pu par une simple lecture des textes se révéler avec autant de netteté. Certes, cela demande un travail préparatoire contraignant, c'est le prix à payer mais l'on a rien sans rien.

Cela étant, il serait possible de faire l'objection suivante : Pourquoi choisir 12 mots et ne pas en prendre 5 ou 10 voire 15 ou 20 ? En fait, la rigueur voudrait que tous soient pris en compte. S'il y a quelques années seuls des programmes informatiques à quelques

milliers d'euros, tels Alceste, pouvaient assumer cette tâche, aujourd'hui les logiciels libres, tel « IRaMuTeQ», sont disponibles et permettent de la réaliser. Toutefois, nous ne saurions trop engager notre lecteur à utiliser de tels programmes sans s'être aguerri à des techniques plus rudimentaires et sans s'être formé à l'analyse des données (Calcul matriciel, analyse en composantes principales et discriminante, classements hiérachiques ascendants et descendants… pour ne citer que quelques techniques) au risque de manipuler des informations dont il ne connaitrait pas toujours la véritable qualité et la valeur.

Aussi, à notre niveau et avec nos moyens, il faut opérer progressivement avec un nombre de mots réduit et voir comment le schéma évolue. On teste d'abord avec 5 mots, puis avec 7, 9 et ainsi de suite. A un moment donné, le schéma se complexifie de telle sorte que l'on sent très bien que l'on a dépassé la limite et l'on revient donc en arrière. Trop peu de mots risque d'enlever toute pertinence à l'analyse, mais trop de mots crée la confusion. En général, on s'aperçoit très vite du nombre optimal de mots à choisir.

5ème exemple : La géographie des mots

Comme nous le faisions remarquer plus haut, la place d'un mot dans un texte n'est pas sans signification. Or, une analyse fondée sur le décompte des occurrences pour efficace qu'elle soit tend à globaliser les problématiques et peut masquer certains aspects par ailleurs importants. Pour illustrer cette question nous prendrons comme exemple le texte du chapitre 9 de l'ouvrage de Michel Henry, « Paroles du Christ ».

Le texte de ce chapitre est constitué de 4358 mots répartis en 22 paragraphes. Nous avons retenu les 20 premiers vocables par ordre d'importance et nous les avons repérés dans chacun des 22 paragraphes. Cela donne le tableau ci-dessous. En gras taille 10, lorsque le nombre d'occurrences dans le paragraphe est supérieur à 10, en gras taille 8 lorsqu' elles sont supérieures à 5 et inférieures à 11 et en taille 8 non gras pour celles inférieures à 6.

Notons au passage que les fréquences relatives des différents substantifs, verbes et adjectifs renseignent tout de même déjà un peu sur le texte.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Total
vie	1	**6**					1			4		4		**6**	**6**	1	2	**6**	4				41
pouvoir											**15**	5	**6**	**7**			2						35
moi										1		**7**	2	**9**		1	2	1		5		1	29
Dieu	3	1	3		3		3		**7**	1			1					3	1		1		27
monde		5		1							1				1	1			**6**	**9**			24
mal							1	3	5	1					1	2	1		3		**6**	1	24
parole	2				2		**8**	2	3						2					2		1	22
Christ	5	2		1	2	1		1	1				2					1	1	5			22
homme	3		2	1			1		2				1	1	1			1	5			1	19
cœur							4	1	2	2					3	2	1	1	2			1	19
lumière																	1	1	**9**	3	4	1	19
propre		1						2		2		5	1	3	1	1	1						17
parabole		4	1	1	4	4	1		1														16
haine	1								1							1			2	**11**			16
absolue										1		1		1	1		3	3	2				12
Jean	3												2						1	4			10
Verbe								1											**6**	1		1	9
vérité															1	1	1	3	2		1		9
règne	1	1	4		1		1																8
auto-révélation							1			1							1	2	2				7

Sous cette forme, le tableau est assez malaisé à analyser. Il est alors trié, non selon le nombre d'occurrences mais selon l'étendue de chacun des mots dans le texte. L'étendue est la différence entre le numéro de paragraphe de sa dernière apparition et celui de sa première. On trouvera ci-dessous le nouveau tableau

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Total	Etend.
homme	3		2	1			1		2				1	1	1			1	5			1	19	22
parole	2				2		**8**	2	3						2					2		1	22	22
Dieu	3	1	3		3		3		**7**	1			1					3	1		1		27	21
Christ	5	2		1	2	1		1	1				2					1	1	5			22	20
monde		5		1							1				1	1			**6**	**9**			24	19
vie	1	**6**					1			4		4		**6**	**6**	1	2	**6**	4				41	19
propre		1						2		2		5	1	3	1	1	1						17	16

cœur						4	1	2	2					3	2	1	1	2			1	19	16
mal						1	3	5	1					1	2	1		3		**6**	1	24	16
moi									1		**7**	2	**9**		1	2	1		5		1	29	13
absolue									1		1		1	1		3	3	2				12	10
parabole		4	1	1	4	4	1		1													16	8
haine	1								1						1			2	**11**			16	7
pouvoir										**15**	5	**6**	**7**			2						35	7
Lumière																1	1	**9**	3	4	1	19	6

On y distingue trois types de mots :

Les mots qui traversent le texte : homme, parole, Dieu, Christ, monde, vie

Ceux qui se répartissent de manière non régulière : cœur, mal, moi, absolue

Enfin, ceux qui se déploient dans une partie spécifique : parabole, pouvoir, lumière, haine

C'est, en général, cette dernière catégorie qui est la plus discriminante et permet de distinguer trois parties principales :

1er partie, de 2 à 10, dans laquelle domine le mot parabole

2ème partie, de 11 à 14, dans laquelle domine le vocable de pouvoir

3ème partie, de 17 à 21, dans laquelle domine les mots de lumière et de haine.

Il suffit alors de vérifier les mots auxquels ces trois vocables sont associés

1° partie : parabole avec parole, Dieu, mal, cœur, Christ et vie

2° partie : pouvoir avec moi et vie

3° partie : lumière et haine avec mal, monde et vie

Restent les paragraphes 15 et 16 qu'il y a lieu de lire afin de bien déterminer s'ils doivent être considérés comme une partie indépendante ou comme appartenant au bloc précédent ou suivant.

La trame est donnée, les mots clés mis en évidence. Il ne suffit plus dès lors que de lire… et relire et re-relire le texte…

Conclusion

Nous voici au terme de notre présentation de quelques techniques de base relatives à la lexicomètrie. Comme nous l'avons dit, il ne s'agissait pas d'en faire une présentation détaillée mais de bien plus modestement permettre d'en mesurer l'intérêt pour l'étude des textes sans devoir posséder de compétences particulières en mathématique et en informatique. Nous espérons avoir répondu à notre projet. Pour ceux qui voudraient aller plus loin, il existe une vaste bibliographie tant sur la lexicomètrie que l'analyse du discours. Il suffit de taper l'un ou l'autre de ces mots dans la barre de recherche de son ordinateur pour en mesurer l'importance. Pour aller encore un peu plus loin, on pourra faire cette même recherche mais dans « Google scholar » où l'on trouvera de nombreux articles tant en français qu'en anglais.