Semaine Data-SHS (7-11 décembre 2020)
"Traiter et analyser des données en sciences humaines et sociales"

Plateforme Universitaire de Données (TIR-PROGEDO)
Université de Grenoble-Alpes
Maison des sciences de l'homme

Frédéric Urien
Laboratoire informatique de Grenoble – Université Grenoble-Alpes
fredericurien@gmail.com

Cyril Labbé
Laboratoire informatique de Grenoble – Université Grenoble-Alpes
cyril.labbe@imag.fr

Dominique Labbé
PACTE (Institut d'Etudes Politiques) - Université Grenoble-Alpes
dominique.labbe@umrpacte.fr

La bibliothèque en ligne du français moderne. L'exemple du vocabulaire de V. Hugo

(mardi 8 décembre 2020)

Une partie de la Bibliothèque électronique du français moderne (XVIIe-XXIe siècle) est en ligne dans une version préliminaire (https://www.lexicometrie.imag.fr). Dans sa version finale, cette bibliothèque comportera 60 millions de mots. A chacun de ces mots est associée une série d'étiquettes comprenant - outre auteur, titre, date, etc. – leur entrée de dictionnaire comportant la graphie standard, le lemme et sa catégorie grammaticale. Pour illustrer l'intérêt de cet outil en sciences humaines, nous présentons quelques aperçus concernant le vocabulaire de V. Hugo.

La semaine Data-SHS organisée par la Maison des Sciences de l'Homme Alpes, dans le cadre des activités de la TGIR Progedo, est l'occasion de présenter la Bibliothèque Électronique du Français Moderne (BEFM) qui vient d'être mise en ligne dans une version de pré-configuration (https://www.lexicometrie.imag.fr).

Dans le domaine des Sciences Humaines et Sociales, une partie significative des données qui doivent être analysées prennent la forme de textes ou de discours pour lesquels des méthodes, des outils, des normes permettent de fournir aux chercheurs des données fiables. La production de ces données, effectuée avec des protocoles reproductibles, constitue le cœur de la lexicométrie.

Après avoir présenté ces outils, nous mettrons l'accent sur deux intérêts de la BEFM. D'une part, la possibilité d'étudier le vocabulaire d'une œuvre ou d'un auteur et non plus simplement des extraits comme dans les corpus habituellement disponibles, ce qui donne l'assurance de traiter l'intégralité du vocabulaire. D'autre part, grâce à certains outils de la lexicométrie, il est possible de connaître les caractéristiques particulières d'un auteur, d'un genre d'une époque… Nous illustrerons ces deux aspects à l'aide de V. Hugo dont on fête le bicentenaire de la première oeuvre (1820). Le vocabulaire d'Hugo ayant déjà été étudié avec les méthodes de la "textométrie" (Brunet 1988 a,b,c), cela permettra au lecteur d'apprécier les limites et les mérites respectifs des différentes approches.

I. PRINCIPALES FONCTIONNALITES DE LA BEFM

La BEFM ne donne pas simplement accès aux mots contenus dans une série de textes mais aussi à une série d'informations produites par des outils de lexicométrie qui seront présentés avant d'aborder la manière de consulter la bibliothèque.

La lexicométrie

La première étape consiste à repérer, dans le texte ou dans la retranscription d'un enregistrement audio, les éléments qui véhiculent le sens. Ce travail peut être réalisé sur un ensemble cohérent de documents constituant un corpus, comme celui des œuvres de V. Hugo.

La première unité est le **mot**, emplacement du texte où apparait une forme graphique. Par exemple, « parce qu'» est un seul mot (une unité de sens insécable malgré le blanc au milieu) et une forme différente de « parce que », les deux étant des variations graphiques d'une même conjonction. A l'inverse, l'article « du » occupe deux mots parce que tous les dictionnaires de

langue sont d'accord pour l'analyser comme la forme contractée de deux unités de sens (de + le).

Il est donc important de rattacher chaque mot à une ou plusieurs unités de sens (vocables). Un **vocable** peut être représenté par une entrée de dictionnaire, composée d'une forme « vedette » (appelée aussi lemme) et d'une catégorie grammaticale. Par exemple, la forme "vedette" retenue pour un verbe est son infinitif, le masculin singulier pour un adjectif, le singulier pour le nom.

Cela conduit à introduire la notion de **forme** (graphique) qui est la flexion sous laquelle un vocable apparaît dans un emplacement du texte. Par exemple, "est" peut être l'apparition du verbe *être* ou du substantif masculin (point cardinal).

Le vocabulaire d'un texte ou d'un corpus est l'ensemble des vocables qui le constituent.

Le lexique d'une langue est l'ensemble des vocables que la langue met à la disposition des locuteurs à un moment donné. Le lexique d'un auteur est l'ensemble des vocables dont il dispose. Ces lexiques ne sont donc pas observables directement. La meilleure représentation possible est fournie par les dictionnaires de langue. Une approche réaliste se fera à l'aide d'échantillons de vocabulaire comme ceux qu'offre la BEFM.

La norme de dépouillement utilisée pour passer des mots aux formes graphiques et aux vocables s'inscrit dans la ligne des travaux de Muller (1963 et 1967 ; Labbé 1990). Elle se calque sur les principes de la lexicographie française et de la pratique des usagers du français.

En suivant ces normes de dépouillement, à l'heure où ces lignes sont écrites, le corpus Hugo de la BEFM contient un peu plus de 2 millions de mots, 63 556 formes graphiques standardisées et un vocabulaire composé de 32 665 vocables différents (tableau en annexe 1).

En approfondissant cette exploration, on remarque que, dans tout texte en français d'une certaine longueur, le verbe *être* est le plus fréquent. Dans le corpus Hugo, il apparaît 62 035 fois (dont 3 080 infinitifs *être*) ; le substantif masculin *être* est utilisé 575 fois, dont 354 fois au masculin singulier. En effet, l'"être" (*humain*) est un thème majeur de V. Hugo qui serait bien difficile à déceler si l'on devait le rechercher parmi les infinitifs homographes (raison pour laquelle Brunet passe à côté). De même, le substantif "est" apparaît 55 fois alors que son homographe, verbe *être* à la troisième personne de l'indicatif présent est utilisé 25 318 fois. Le substantif "été" : 153 fois contre 1 945 fois pour le participe passé. Pourquoi faudrait-il perdre le point cardinal ou la saison favorite d'Hugo ? Sans compter la conjonction *soit*, les *sommes* (d'argent ou court sommeil ?), le *fût*, les *étais*…

Pour résoudre ces difficultés, à chaque mot est attachée une étiquette comportant sa forme graphique standardisée, son lemme et sa catégorie grammaticale. Cette opération s'appelle annotation, étiquetage (*tagging*) ou lemmatisation des textes.

Cyril Labbé et Dominique Labbé ont développé des outils informatiques qui permettent de réaliser cette étiquetage de manière automatisée pour la plus grande partie du traitement, avec le souci de limiter autant que possible le nombre d'erreurs (par rapport à la norme de dépouillement : pour les textes mis en ligne ce taux est inférieur à 0,5%).

L'ensemble des textes ainsi annotés constitue la BEFM, soit un corpus de plus de 60 millions de mots qui, une fois en ligne, permettra aux chercheurs, enseignants et étudiants ou simples curieux d'analyser un auteur, une œuvre, une époque en suivant une démarche comme celle qui est présentée dans la seconde partie de ce texte.

Début décembre 2020, seule une partie limitée de la BEFM a été déployée, l'ensemble du corpus public sera rendu accessible dans les prochains mois.

Le format XML TEI

Le format des fichiers informatiques est un élément essentiel pour le partage, la réutilisation et la pérennité des informations. En matière de fichiers annotés, la Text Encoding Initiative (TEI) fournit un ensemble de spécifications basées sur le format XML. Elle offre de nombreuses possibilités de mise en forme des informations et peut accueillir des extensions. Enfin, les fichiers au format TEI sont exploitables par de nombreux logiciels qui effectuent des traitements sur les fichiers XML, comme par exemple l'interface TXM utilisée pour mettre en ligne la BEFM (présentation dans Heiden, Magué et Pincemin 2010).

Le format TEI propose des outils pour associer au texte des métadonnées (auteurs, titres, année, genre…) qui sont décrites plus bas. D'autres outils organisent le document lui-même. Pour la BEFM, nous avons fait le choix d'utiliser ce format normalisé et de nous focaliser sur la partie qui décrit le texte lui-même. A titre d'exemple, la figure 1 présente une phrase de "Mélite ou les fausses lettres", comédie de Pierre Corneille encodée au format XML TEI.

Chaque mot occupe une ligne avec, dans l'ordre, le lemme, la catégorie grammaticale - décomposée en "pos" (part of speech) et "mds" (morpho-syntactic description) – son rang dans le corpus, enfin : le mot (w) du texte dans sa graphie standard. Les "pos" et "mds" sont décrites plus bas dans le paragraphe consacré aux requêtes.

Figure 1. Phrase annotée issue de "Mélite", comédie de Pierre Corneille

```
<s>
        <w lemma="mon" pos="DET" msd="pos" id="18782">ma</w>
        <w lemma="soeur" pos="NOM" msd="fem" id="18783">soeur</w>
        <w lemma="," pos="PON" msd="fbl" id="18784">,</w>
        <w lemma="acquitter" pos="VER" msd="pst" id="18785">acquitte</w>
        <w lemma="moi" pos="PRO" msd="per" id="18786">moi</w>
        <w lemma="de" pos="PRE" id="18787">d'</w>
        <w lemma="un" pos="DET" msd="art" id="18788">une</w>
        <w lemma="reconnaissance" pos="NOM" msd="fem" id="18789">reconnoissance</w>
        <w lemma="dont" pos="PRO" msd="aut" id="18790">dont</w>
        <w lemma="un" pos="DET" msd="art" id="18791">un</w>
        <w lemma="autre" pos="DET" msd="ind" id="18792">autre</w>
        <w lemma="destin" pos="NOM" msd="mas" id="18793">destin</w>
        <w lemma="je" pos="PRO" msd="per" id="18794">m'</w>
        <w lemma="avoir" pos="VER" msd="pst" id="18795">a</w>
        <w lemma="mettre" pos="VER" msd="ppa" id="18796">mise</w>
        <w lemma="en" pos="PRE" id="18797">en</w>
        <w lemma="impuissance" pos="NOM" msd="fem" id="18798">impuissance</w>
        <w lemma=":" pos="PON" msd="fbl" id="18799">:</w>
        <w lemma="accorder" pos="VER" msd="pst" id="18800">accorde</w>
        <w lemma="ce" pos="DET" msd="dem" id="18801">cette</w>
        <w lemma="grâce" pos="NOM" msd="fem" id="18802">grâce</w>
        <w lemma="à" pos="PRE" id="18803">à</w>
        <w lemma="notre" pos="DET" msd="pos" id="18804">nos</w>
        <w lemma="juste" pos="ADJ" msd="pur" id="18805">justes</w>
        <w lemma="désir" pos="NOM" msd="mas" id="18806">désirs</w>
        <w lemma="." pos="PON" msd="frt" id="18807">.</w>
```

L'application Textométrie (TXM)

Pour visualiser ces données, il faut une interface. Notre choix s'est porté sur l'application TXM qui est développée par l'ENS de Lyon depuis 2007 (Haiden & Al. 2010) (Figure 2). Elle fonctionne sur un ordinateur de bureau ou sur un portail accessible en ligne. Elle bénéficie de mises à jour régulières, d'une équipe de développement qui répond rapidement aux sollicitations et elle permet de réaliser des analyses sur les corpus au format XML TEI.

Figure 2. Page d'accueil du site Lexicométrie

Le projet BEFM est porté par l'équipe Sigma du Laboratoire d'Informatique de Grenoble. Pour l'instant, il y a en ligne, dans une version préliminaire, un peu plus de 10% de la

bibliothèque totale. En 2021, le service devrait monter en puissance, intégrer tous les textes du domaine public présents dans la bibliothèque et mettre à disposition des outils supplémentaires.

Pour commencer l'exploration il convient de faire un clic droit sur le nom du corpus (BEFM). On dispose alors d'une série de fonctionnalités et de requêtes possibles.

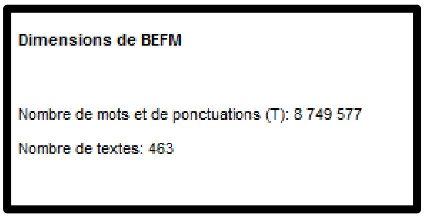
1. *Principales fonctionnalités.*

Les principales fonctionnalités disponibles sont les suivantes :

<u>Textes :</u> affiche la liste des textes ainsi que leurs métadonnées.

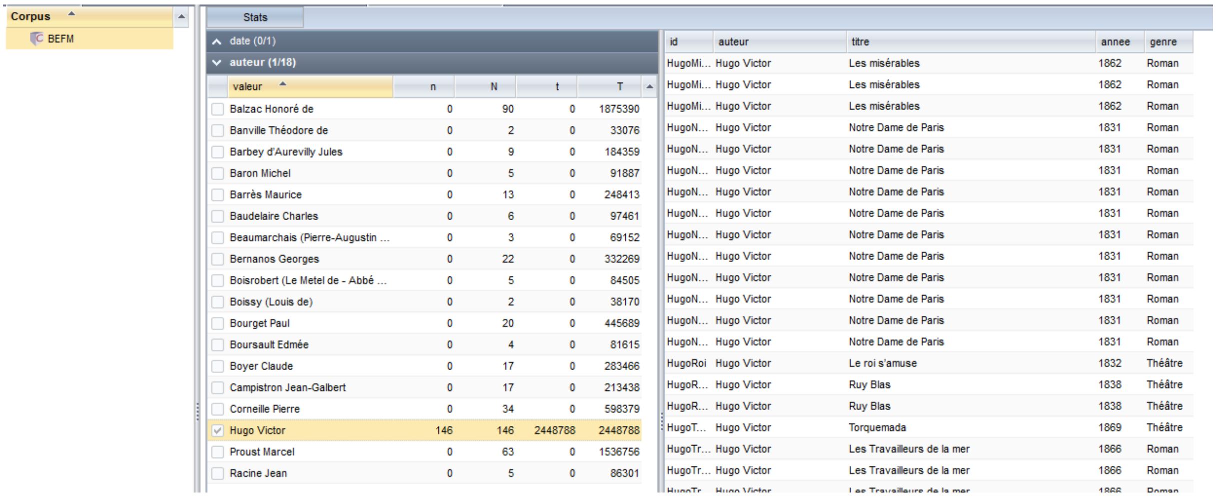
<u>Dimensions :</u> affiche les dimensions du corpus (figure 3).

Figure 3. Dimensions de la version préliminaire de la BEFM (décembre 2020)



<u>Créer un sous-corpus - textes :</u> permet de constituer un corpus de travail contenant une partie des textes disponibles dans la BEFM. Le choix peut être fait à partir des métadonnées associées aux textes ou directement à partir des textes eux-mêmes (figures 4, 5 et annexe 2).

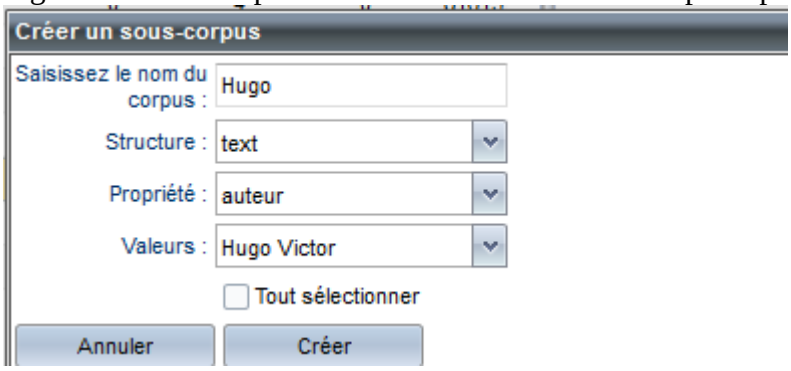
Figure 4. Eléments permettant de constituer un sous-corpus à partir des textes



valeur	n	N	t	T
Balzac Honoré de	0	90	0	1875390
Banville Théodore de	0	2	0	33076
Barbey d'Aurevilly Jules	0	9	0	184359
Baron Michel	0	5	0	91887
Barrès Maurice	0	13	0	248413
Baudelaire Charles	0	6	0	97461
Beaumarchais (Pierre-Augustin ...	0	3	0	69152
Bernanos Georges	0	22	0	332269
Boisrobert (Le Metel de - Abbé ...	0	5	0	84505
Boissy (Louis de)	0	2	0	38170
Bourget Paul	0	20	0	445689
Boursault Edmée	0	4	0	81615
Boyer Claude	0	17	0	283466
Campistron Jean-Galbert	0	17	0	213438
Corneille Pierre	0	34	0	598379
Hugo Victor	146	146	2448788	2448788
Proust Marcel	0	63	0	1536756
Racine Jean	0	5	0	86301

id	auteur	titre	annee	genre
HugoMi...	Hugo Victor	Les misérables	1862	Roman
HugoMi...	Hugo Victor	Les misérables	1862	Roman
HugoMi...	Hugo Victor	Les misérables	1862	Roman
HugoN...	Hugo Victor	Notre Dame de Paris	1831	Roman
HugoN...	Hugo Victor	Notre Dame de Paris	1831	Roman
HugoN...	Hugo Victor	Notre Dame de Paris	1831	Roman
HugoN...	Hugo Victor	Notre Dame de Paris	1831	Roman
HugoN...	Hugo Victor	Notre Dame de Paris	1831	Roman
HugoN...	Hugo Victor	Notre Dame de Paris	1831	Roman
HugoN...	Hugo Victor	Notre Dame de Paris	1831	Roman
HugoN...	Hugo Victor	Notre Dame de Paris	1831	Roman
HugoN...	Hugo Victor	Notre Dame de Paris	1831	Roman
HugoN...	Hugo Victor	Notre Dame de Paris	1831	Roman
HugoN...	Hugo Victor	Notre Dame de Paris	1831	Roman
HugoRoi	Hugo Victor	Le roi s'amuse	1832	Théâtre
HugoR...	Hugo Victor	Ruy Blas	1838	Théâtre
HugoR...	Hugo Victor	Ruy Blas	1838	Théâtre
HugoT...	Hugo Victor	Torquemada	1869	Théâtre
HugoTr...	Hugo Victor	Les Travailleurs de la mer	1866	Roman
HugoTr...	Hugo Victor	Les Travailleurs de la mer	1866	Roman

Figure 5. Eléments permettant de réaliser un sous corpus à partir des métadonnées



Créer une partition : permet de former un ensemble de données basé sur les métadonnées. Par exemple, pour un auteur, il est possible d'avoir le nombre d'occurrences des mots sur l'ensemble de ses textes (dans la figure 6, pour afficher le nom complet des auteurs, élargir la colonne).

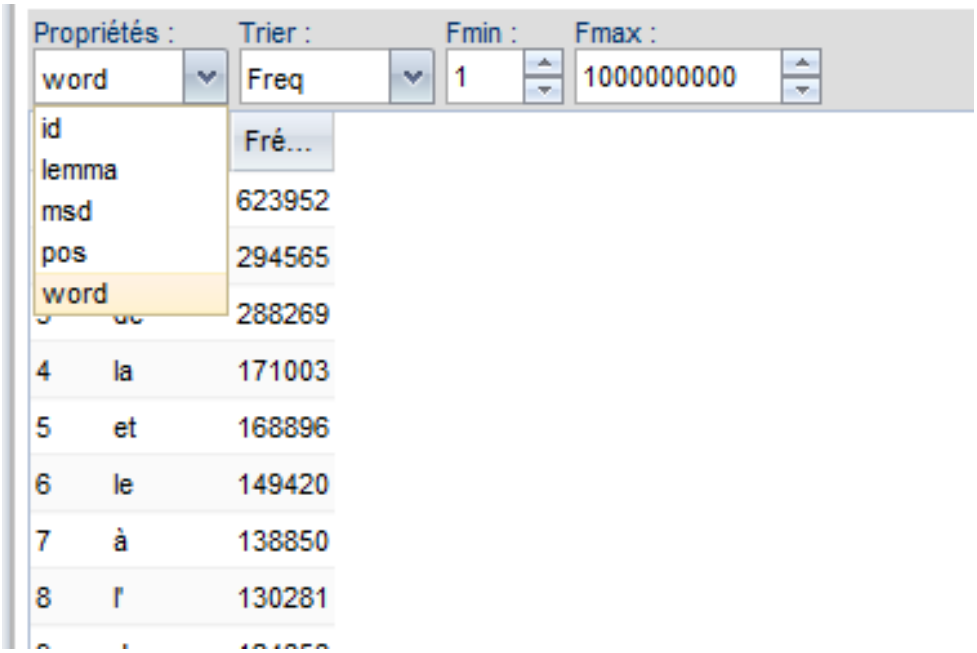
Figure 6. Table hiérarchique des formes graphiques de la BEFM ventilées par auteurs classés par ordre alphabétique (colonnes).

	P...	Bal...	B...	Ba...	B...	Ba...	B...	B...	Be...	B...	B...	Bo...	B...	Bo...	Ca...	Co...	Hug...	Pro...	R...
1	,	131329	3244	15457	7606	17688	7657	4331	25120	7635	2600	34115	5252	20813	15766	37773	174577	107295	5694
2	.	60830	941	3815	4020	8128	2182	2720	12107	2244	1933	12447	2922	8600	8166	16124	108899	35408	3079
3	de	68494	838	7231	1991	9598	3429	1519	10641	2052	753	16341	2059	7241	5432	15652	74255	58576	2167
4	la	41201	653	3966	1071	5354	2039	997	6388	1001	566	8965	916	3414	2557	6789	54264	29935	927
5	et	31129	1027	4128	1137	5107	2685	867	4479	1959	788	8236	1512	6570	3915	14440	52991	26245	1681
6	le	35740	612	3197	1436	4275	1747	1086	5720	1058	646	6547	1278	4203	3017	7980	47415	22206	1257
7	à	31160	251	2784	1241	3878	1343	841	5092	983	499	7323	1288	3717	3100	9500	37017	27656	1177
8	l'	24042	505	2703	999	3887	1852	919	5320	1092	617	6701	1225	3866	2827	9122	43330	20048	1226
9	des	31201	858	2502	482	5580	2110	542	3592	388	216	5432	760	2390	2352	5169	37114	23066	902
10	il	22986	224	2567	1447	4027	948	1037	5424	1213	527	6864	1023	3123	2232	7950	34097	18593	983
11	un	22046	406	1847	787	3024	1449	769	4546	665	456	4888	1086	4057	2395	7683	27765	17893	1004
12	que	15858	262	1946	1720	2040	923	1031	3690	1579	461	4761	1360	4298	3487	9960	22727	24793	1311
13	les	27427	631	2218	437	4048	1522	423	2795	400	212	3993	524	1424	1915	3800	30231	16892	683
14	d'	19104	332	1953	551	2754	1187	540	3941	811	300	5648	813	3219	1985	6748	21574	19888	848
15	du	20505	320	2038	440	2940	1104	500	3334	402	250	4486	626	1960	1396	4101	29980	15968	638

2. Principales informations

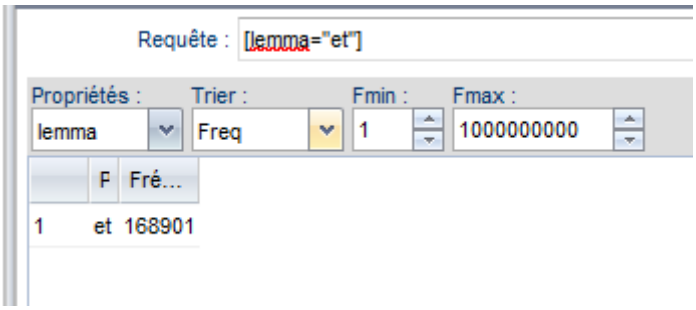
Lexique : permet d'afficher le nombre d'occurrences – dans toute la bibliothèque ou dans un sous-corpus - pour les propriétés associées à chaque mot : lemma, word, pos, msd (la figure 7 donne la table hiérarchique des formes graphiques de la BEFM).

Figure 7. Table hiérarchique des formes (word) figurant dans le corpus de la BEFM



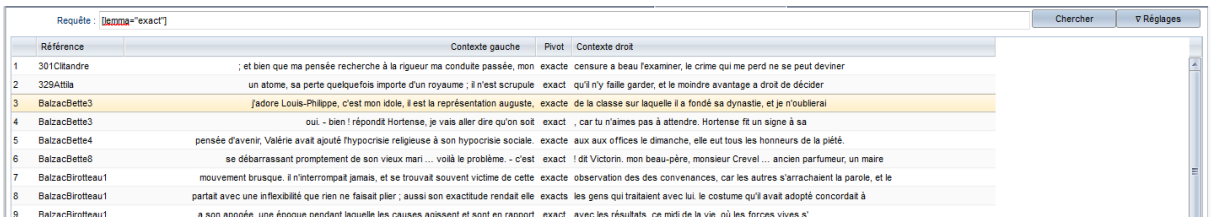
Index : permet d'obtenir les effectifs d'un mot, lemme ou catégorie grammaticale dans toute la bibliothèque ou le sous-corpus sous revue (figure 8).

Figure 8. Fréquence de parution du lemme "et" dans le corpus de la BEFM



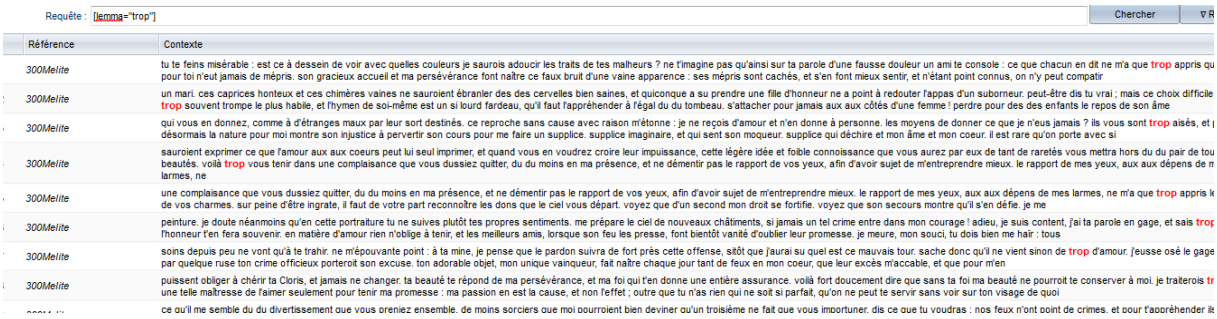
Concordance : retourne un extrait de chaque texte comportant l'élément correspondant à la requête comme pivot ainsi que 15 mots avant et après (figure 9).

Figure 9. Concordance du lemme *exact* dans la BEFM



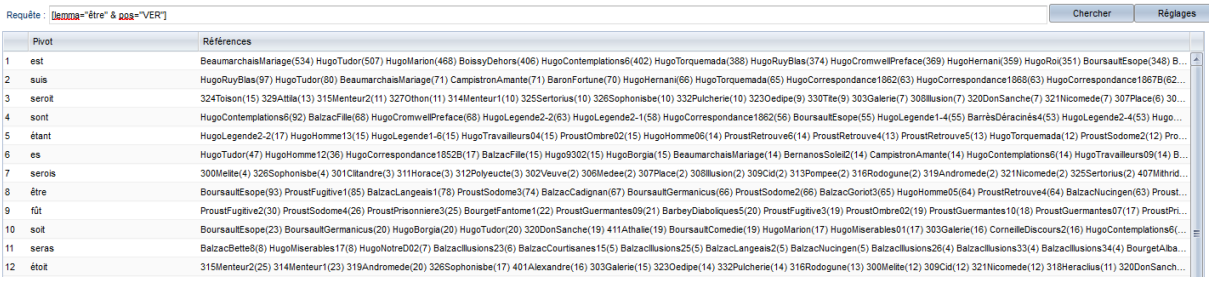
Contexte : affiche le mot correspondant à la requête, en rouge, dans son contexte plus large que la concordance mais sans alignement (figure 10).

Figure 10. Contexte du lemme *trop* dans la BEFM



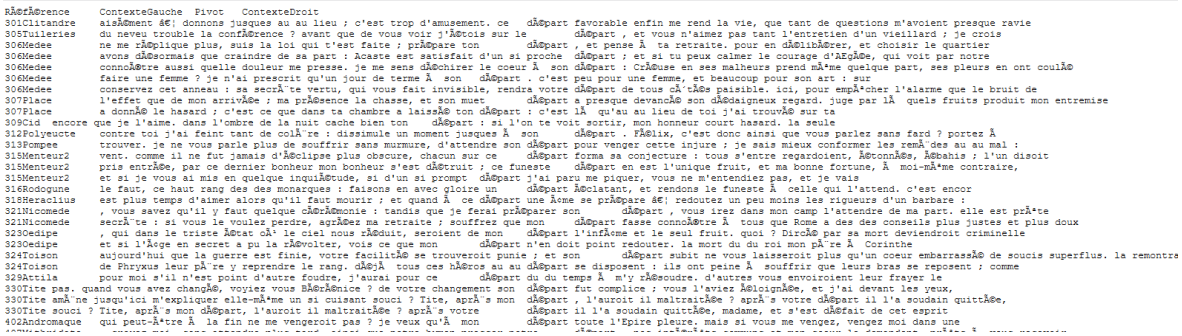
Références : la liste des textes qui contiennent l'élément recherché, ainsi que le nombre d'occurrence dans chacun des textes (figure 11).

Figure 11. Nombre d'occurrences du mot revoir dans chaque texte du corpus



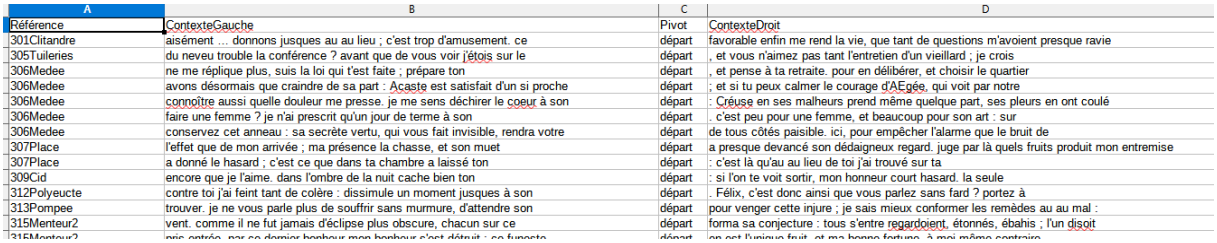
Exporter : permet d'exporter le résultat de la requête au format txt (figure 12 et 13)

Figure 12. Fichier txt mis à disposition par l'application TXM pour exporter le résultat d'une requête.



Renommé en csv. Le fichier est utilisable dans une feuille de tableur avec une tabulation comme séparateur et UTF 8 comme encodage (figure 13).

Figure 13 : le même fichier dont l'extension a été changée en ;csv, ouvert sous un tableur.



	A	B	C	D
	Référence	ContexteGauche	Pivot	ContexteDroit
	301Clitandre	aisément … donnons jusques au au lieu ; c'est trop d'amusement. ce	départ	favorable enfin me rend la vie, que tant de questions m'avoient presque ravie
	305Tuileries	du neveu trouble la conférence ? avant que de vous voir j'étois sur le	départ	, et vous n'aimez pas tant l'entretien d'un vieillard ; je crois
	306Medee	ne me réplique plus, suis la loi qui t'est faite ; prépare ton	départ	, et pense à ta retraite. pour en délibérer, et choisir le quartier
	306Medee	avons désormais que craindre de sa part : Acaste est satisfait d'un si proche	départ	; et si tu peux calmer le courage d'AEgée, qui voit par notre
	306Medee	connoître aussi quelle douleur me presse. je me sens déchirer le coeur à son	départ	: Créuse en ses malheurs prend même quelque part, ses pleurs en ont coulé
	306Medee	faire une femme ? je n'ai prescrit qu'un jour de terme à son	départ	. c'est peu pour une femme, et beaucoup pour son art : sur
	306Medee	conservez cet anneau : sa secrète vertu, qui vous fait invisible, rendra votre	départ	de tous côtés paisible. ici, pour empêcher l'alarme que le bruit de
	307Place	l'effet que de mon arrivée ; ma présence la chasse, et son muet	départ	a presque devancé son dédaigneux regard. juge par là quels fruits produit mon entremise
	307Place	a donné le hasard ; c'est ce que dans ta chambre a laissé ton	départ	: c'est là qu'au au lieu de toi j'ai trouvé sur ta
	309Cid	encore que je l'aime. dans l'ombre de la nuit cache bien ton	départ	: si l'on te voit sortir, mon honneur court hasard. la seule
	312Polyeucte	contre toi j'ai feint tant de colère : dissimule un moment jusques à son	départ	. Félix, c'est donc ainsi que vous parlez sans fard ? portez à
	313Pompee	trouver. je ne vous parle plus de souffrir sans murmure, d'attendre son	départ	pour venger cette injure ; je sais mieux conformer les remèdes au au mal :
	315Menteur2	vent. comme il ne fut jamais d'éclipse plus obscure, chacun sur ce	départ	forma sa conjecture : tous s'entre regardoient, étonnés, ébahis ; l'un disoit
	315Menteur2	pris entrée, par ce dernier bonheur mon bonheur s'est détruit : ce funeste	départ	en est l'unique fruit, et ma bonne fortune, à moi même contraire

Les interrogations sont réalisées grâce à CQL (Corpus Query Language), qui est un langage de requête – mis au point dans les années 1990 pour le Open Corpus Workbench pour le British National Corpus - utilisé maintenant par de très nombreux outils de "fouille de textes".

Les requêtes CQL

Les requêtes peuvent porter sur les différentes propriétés qui sont attachées aux mots. La BEFM en offre 4 : word, lemma, pos et msd.

word : la forme graphique standardisée du texte d'origine

lemma : l'entrée du dictionnaire associée à cette forme. Une recherche sur [lemma="aimer" & pos ="VER"] donne les occurrences du lemme *aimer* avec ses flexions (aime, aimera, aimeront, aimant, aimer…) à l'exclusion du substantif *aimant*.

pos : catégorie grammaticale du mot : adverbe, nom, verbe, adjectif… Tous les mots de la BEFM disposent d'une balise pos utilisable dans une requête - à l'exclusion des lettres euphoniques comme "t" (dans "que cherche-t-on ?") ou "l'" dans "l'on est bien avancé !". "pos" peut prendre les valeurs suivantes :

ADJ adjectifs
ADV adverbes
CON conjonctions
DET pour les déterminants
ETR mots étrangers
INT interjections
NOM noms
PON ponctuations
PRE prépositions
PRO pronoms
VER verbes

msd : fournit des informations morphosyntaxiques complémentaires sur le mot. Il permet, si on le souhaite, de limiter la recherche à une sous-catégorie grammaticale. Par exemple la recherche [pos="VER"] retourne tous les verbes alors que [msd="fut"] retournera seulement les verbes au futur. msd peut prendre les valeurs suivantes :

pur : adjectif “pur”
ppa : participe passé dans un emploi “adjectivé”
coo : conjonction de coordination
sub : conjonction de subordination
art : article (défini ou indéfini)
num : déterminant numéral ou cardinal
pos : déterminant possessif
dem : déterminant démonstratif
ind : déterminant indéfini
pro : nom propre
mas : nom commun masculin
fem : nom commun féminin
fbl : ponctuation faible (interne à la phrase)
frt : ponctuation forte (délimitant la phrase)
per : pronom personnel
aut : autres pronoms (relatif, interrogatif, possessif, etc.)
ppa : verbe participe passé
ppe : verbe participe présent
inf : verbe à l’infinitif
con : verbe au conditionnel
pst : verbe au présent (indicatif, subjonctif, impératif)
ipf : verbe à l’imparfait
psp : verbe au passé simple

A partir de ces éléments, il est possible de construire une requête en utilisant le formalisme propre à CQL. Concernant ce formalisme, de nombreuses ressources sont disponibles sur internet (voir la sitographie dans les références). Quelques exemples sont proposés ci-dessous afin de commencer l’exploration de la BEFM. D’autres sont fournis par l’aide en ligne de la BEFM

Pour rechercher le vocable "manger, nom masculin" la requête sera :
[word="manger" & msd="mas"]
Les occurrences de "manger, verbe à l’infinitif" : [word="manger" & msd="inf"]
Celles de "suivre, verbe avec la forme suis" : [word="suis" & lemma="suivre"]
Le pronom “je” suivi du verbe "manger" dans toutes les flexions possibles :
[word="je"][lemma="manger" & pos="VER”]
Le nom masculin singulier “être” suivi de l’adjectif "cher” :
[word="être" & pos="NOM"][word="cher" & pos="ADJ"]
Etc.

La bibliothèque permet d’étudier certains aspects peu connus de la langue française. Par exemple, la "modalisation verbale", consistant à placer devant un verbe à l’infinitif un autre verbe jouant le rôle d’auxiliaire "modalisateur" (Benveniste 1965 & 1970 ; Labbé & Labbé 2010). Voici des requêtes possibles :

[pos="VER"][msd="inf"] (vouloir dire) ou

[pos="VER"][word=".*"][msd="inf"] (le deuxième terme indique qu'un mot quelconque peut s'insérer entre les deux verbes : veut-elle dire ? ne veut pas dire, vouloir bien dire, etc.)

Autre exemple de requête complexe :

[pos="NOM"][pos="PRE"][pos="DET"][pos="NOM"]

retourne toutes les constructions comme "nom de la mère", "président de la république", etc.

Si l'on souhaite faire ces requêtes sur une partie seulement de la BEFM, une étape préalable consiste à réaliser un sous-corpus comme indiqué ci-dessus et dans l'annexe 2. Cette opération s'appuie sur les métadonnées fournies à TXM en même temps que les fichiers TEI-XML.

Métadonnées

Ces métadonnées découlent des informations disponibles dans la BEFM. La liste présentée dans le tableau 1 augmentera au fur et à mesure que la bibliothèque grandira. Ces métadonnées permettent de constituer des sous-corpus variés, de manière à répondre à un grand nombre de questionnements.

Tableau 1. Métadonnées disponibles (décembre 2020), statut et valeurs possibles

Champ CQL	Statut	Valeurs possibles
id	Pour tous les textes	Nom du fichier sans extension
section	Pour tous les textes	Littérature Oral Politique Presse
auteur	Pour tous les textes	Nom Prénom
titre	Lorsque disponible	Libre
soustitre	Lorsque disponible	Libre
annee	Pour tous les textes	AAAA
genre	Pour tous les textes	Allocution (discours politique) Conférence de presse (discours politique) Correspondance (littérature) Entretien (discours politique) Message (discours politique) Poésie (littérature) Roman (littérature) Théâtre (littérature)

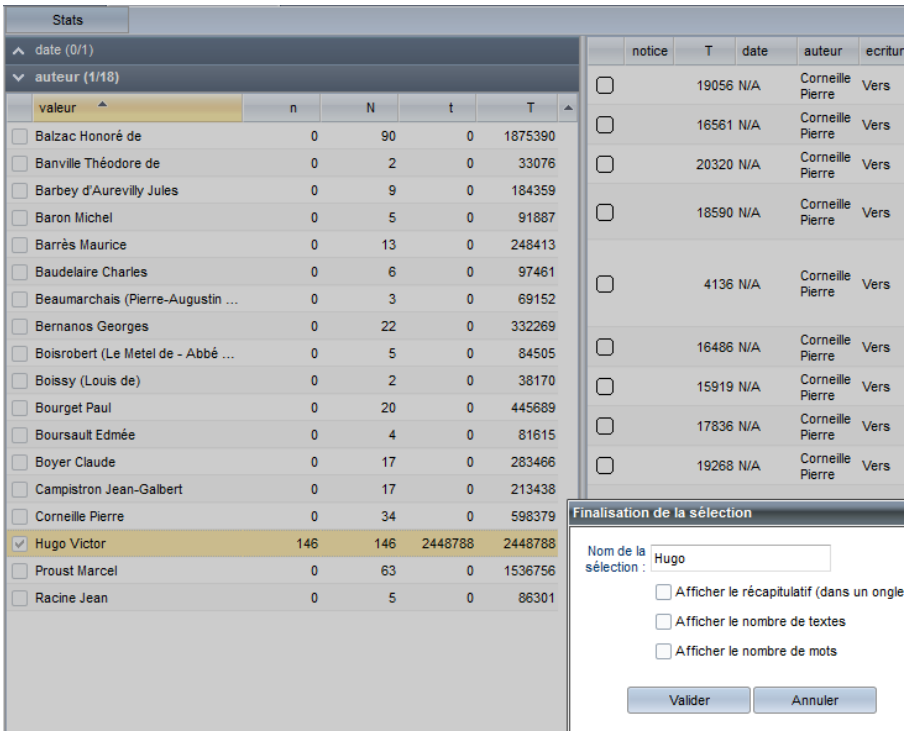
sousgenre	Lorsque disponible	Comédie (théâtre) Essai (littérature) Nouvelle (littérature) Radio (discours politique) Télévision (discours politique) Tragédie (théâtre) Tragi-comédie (théâtre)
ecriture	Pour tous les textes	Prose Vers
date	Lorsque disponible	JJ MMMMMM AAAA
role	Lorsque disponible	Candidat (discours politique) Chef d'état (discours politique) Chef de gouvernement (discours politique)
pays	Pour tous les textes	Pays de l'auteur

Pour certains attributs, plusieurs valeurs sont possibles : ainsi un texte peut être écrit sur plusieurs années, peut être en partie en prose et en vers, une allocution peut avoir été diffusée à la radio et à la télévision, un document peut avoir plusieurs auteurs. Les valeurs multiples sont séparées par des tirets (-) et le langage de requête CQL permet de faire une requête sur une partie seulement de la valeur d'un attribut en indiquant que la valeur recherchée peut avoir un préfixe ou un suffixe. La formulation pourrait être la suivante : sousgenre=".*Radio*" pour indiquer que le sous-genre doit contenir le mot Radio avec éventuellement d'autres indications avant ou après (astérisques).

Une fois familiarisé avec le langage CQL et la lexicométrie, quels résultats peut-on espérer obtenir ? La recherche des attestations de certains mots ou des combinaisons de mots est évidente. D'autres utilisations plus élaborées sont possibles. La suite de cette communication donne un exemple tiré du corpus Victor Hugo (voir figure 14) qui est quasiment complet dans la version préliminaire de la BEFM[1].

[1] Il manque les *Odes et ballades*, *l'Art d'être grand-père* et le *Rhin* qui seront intégrés bientôt dans la bibliothèque. Les autres textes sont en cours de correction.

Figure 14 : Constitution du sous-corpus Victor Hugo : faire un clic droit sur BEFM puis Sous-corpus > texte > ouvrir l'onglet auteur > cocher Hugo Victor > Valider > Donner un nom et cocher les cases > Valider



II. LE VOCABULAIRE D'HUGO

Selon un postulat habituel, les vocables les plus utilisés traduisent les préférences d'un auteur.

Le tableau 2 apporte une première réponse limitée aux 40 premiers vocables rangés par ordre décroissant d'effectifs[2]. La colonne F (pour fréquence) donne la densité relative moyenne du vocable dans le corpus. Cette fréquence est obtenue en rapportant l'effectif du vocable (colonne N) au nombre total de mots compris dans le corpus (une fois les ponctuations déduites, dernière ligne du tableau en annexe). Par exemple, Hugo utilise le déterminant "le" en moyenne 100,65 fois pour mille mots (210 677 / 2 093 177 *1 000). Si la convention « pour mille mots » s'est imposée à la place des pourcentages, cela tient à ce que ces densités sont très faibles. Ici dès le 14e vocable, cette fréquence passe en dessous de 1%.

On s'attend à ce que la tête de l'index renseigne sur les mots favoris de l'auteur or, dans les 20 premiers rangs, ne figure aucun substantif ou adjectif et aucun verbe en dehors de *être* –

[2] Toutes les valeurs, données dans la suite de cette communication, évolueront légèrement dans quelques jours quand seront intégrés dans la bibliothèque les textes signalés ci-dessus.

en 3e position - *avoir* (8e) que l'on considère habituellement comme des outils, du moins dans leur emploi usuel d'auxiliaire.

Tableau 2. Les quarante vocables les plus fréquents dans le corpus Hugo (Requête : corpus Hugo > Lexique > Réglage > Propriété = lemma > Calculer)

Rang	Lemme (catégorie grammaticale)	N (occurrences)	F Fréquence en pour mille mots
1	le (dét)	210 677	100,65
2	de (pré)	130 348	62,27
3	être (v)	62 035	29,64
4	et (cj)	53 516	25,57
5	à (pré)	53 005	25,32
6	un (dét)	43 670	20,86
7	il (pro)	42 832	20,46
8	avoir (v)	37 967	18,14
9	je (pro)	33 843	16,17
10	ce (dét)	24 024	11,48
11	que (cj)	24 023	11,48
12	ne (adv)	23 596	11,27
13	se (pro)	21 014	10,04
14	ce (pro)	20 286	9,69
15	son (dét)	19 501	9,32
16	vous (pro)	19 483	9,31
17	qui (pro)	19 339	9,24
18	dans (pré)	18 769	8,97
19	en (pré)	15 738	7,52
20	que (pro)	15 119	7,22
21	le (pro)	14 315	6,84
22	on (pro)	13 077	6,25
23	faire (v)	12 947	6,19
24	pas (adv)	12 056	5,76
25	pour (pré)	10 893	5,20
26	dire (v)	10 054	4,80
27	sur (pré)	10 049	4,80
28	mon (dét)	9 347	4,47
29	plus (adv)	8 680	4,15
30	tout (dét)	8 008	3,83
31	comme (cj)	7 641	3,65
32	y (pro)	7 616	3,64
33	par (pré)	7 574	3,62
34	tu (pro)	7 481	3,57
35	lui (pro)	7 437	3,55
36	nous (pro)	7 194	3,44
37	voir (v)	6 641	3,17
38	homme (n m)	6 295	3,01
39	avec (pré)	6 212	2,97
40	mais (cj)	5 683	2,72

Le tableau 2 montre que, dans tout corpus, les mots les plus fréquents jouent un rôle essentiellement syntaxique et n'acquièrent un contenu qu'avec le contexte de l'énonciation, raison pour laquelle l'habitude s'est prise de les baptiser « outils ». Les premiers mots non-outils sont loin dans la liste. Ainsi, le substantif le plus utilisé (homme) - une particularité d'Hugo – figure en 38e position (il apparaît en moyenne trois fois dans chaque mille mots), l'adjectif grand en 57e position, puis viennent monsieur (58e), jour (65e) et chose (67e), etc.

Les 100 vocables les plus utilisés occupent 60% de toute la surface de l'oeuvre. Les 32 565 suivants se partagent donc les 40% restants. Ce sont pourtant eux qui véhiculent l'essentiel du sens. Parmi ces 100 premiers, 67 (soit les deux tiers) sont des mots outils, baptisés ainsi parce qu'ils ont un rôle essentiellement syntaxique et parce que leur contenu varie en fonction de chaque énoncé.

Ces 67 outils sont utilisés 1,135 millions de fois, soit 54% de la surface du corpus.

Autrement dit, le vocabulaire d'un corpus est une grande collection d'événements rares très inégalement distribués, dont les plus fréquents ne sont pas les plus importants pour le sens. Ce phénomène d'inégale distribution des fréquences a été mis en lumière par Zipf 1935 (voir également Mandelbrot, 1957). Du coup, le sens devient un peu l'aiguille à rechercher dans la botte de foin !

La solution consiste à regrouper les vocables par catégories grammaticales et à convertir les effectifs (absolus) en fréquences (relatives), ce qui permet la comparaison de corpus de longueurs différentes. Par exemple, pour les substantifs, les effectifs sont obtenus de la BEFM à l'aide de la fonction « index » avec la requête [pos="NOM"]. Il suffit d'exporter ces données et de les traiter avec un tableur. Les tableaux 4 à 6 donnent ces renseignements pour les substantifs, les adjectifs et les verbes.

On peut alors répondre à deux questions. D'une part, quels sont les vocables les plus employés dans le corpus Hugo ? D'autre part, qu'est-ce qui singularise le vocabulaire de cet auteur par rapport au reste de la littérature française ? Nous examinerons d'abord cette seconde question.

Comment repérer les singularités d'un auteur ?

Prenons l'exemple des substantifs. Par rapport au reste de la littérature, quels sont ceux que Hugo préfère ou évite ? L'usager de la bibliothèque électronique, dès qu'elle sera complète, pourra répondre assez simplement à cette question en appliquant, aux données fournies par la BEFM, un calcul proposé par Muller 1977 (chapitre 9). A condition que les effectifs absolus soient assez grands (au moins une dizaine d'occurrences chez Hugo) ce calcul est une bonne

approximation de la loi hypergéométrique. Une démarche semblable a été faite par Brunet (1988b) mais l'auteur l'a limitée aux suremplois de quelques substantifs et adjectifs ne posant pas de problèmes d'homographie.

Considérons d'abord le substantif le plus employé par Hugo (*homme*). Dans l'ensemble de la section littéraire de la bibliothèque électronique, les substantifs apparaissent 4 737 652 fois dont 49 561 sont des occurrences du vocable *homme* (tableau 3) et, parmi eux, 6 295 proviennent de Hugo. Il est donc possible de constituer, sur une feuille de calcul, deux ensembles disjoints : Hugo et tous les autres auteurs (dernière colonne du tableau 3).

Tableau 3. Poids des substantifs dans le corpus Hugo

	Hugo	Total littérature	Autres (Littérature sans Hugo)
Total substantifs (*Ns*)	409 392	4 737 652	4 328 260
Homme (n_i)	6 224	49 561	43 337

Combien de fois Hugo aurait-il utilisé ce vocable s'il avait suivi l'usage moyen des autres écrivains (dernière colonne du tableau) ? Cela revient à considérer que les 409 392 substantifs d'Hugo (N_{sh}) sont tirés au sort dans une urne contenant les 4,328 millions de substantifs de tous les autres (N_{sa}). Combien de fois ce tirage donnerait-il *homme* ? Pour obtenir cet effectif théorique ($ntheo_i$), on multiplie le total des substantifs de Hugo (N_{sh}) par la fréquence de *homme* chez les autres (n_{ai}) :

$$\text{ntheo}_i = N_{sh} * \frac{n_{ai}}{N_{sa}} = 409\,392 * \frac{43\,337}{4\,328\,260} = 4\,099\ mots$$

Soit un écart, entre la valeur observée et la valeur attendue, de 6 224 – 4 099 = + 2 125 mots (soit +51% par rapport à la moyenne de tous les autres écrivains présents dans la bibliothèque). Mais peut-on affirmer qu'il s'agit bien d'un trait caractéristique de cet auteur et non d'un effet du hasard ?

Considérons l'hypothèse formulée ci-dessus d'un usage semblable chez Hugo par rapport à tous les autres. Dans ce cas, la probabilité qu'un substantif employé par Hugo soit le vocable *homme* serait de :

$$p = \frac{4\,099}{409\,392} = 0{,}01001$$

Et la probabilité complémentaire (qu'un substantif ne soit pas *homme*) :

$$q = 1 - p = 0{,}99009$$

D'où l'on tire un écart type théorique (ou déviation standard) autour de la valeur attendue :

$$\sigma = \sqrt{Nsh.p.q} = \sqrt{409\,392 * 0{,}01001 * 0{,}99009} = 63{,}7\ mots$$

Et un "écart réduit" (ou z-score), en rapportant à l'écart type, l'écart entre l'effectif constaté et la valeur attendue :

$$z = \frac{2\,125}{63{,}7} = 33{,}4$$

Dans le cas d'une population unique (Hugo et tous les autres utilisent *homme* de la même façon), 66% des observations seraient comprises entre la moyenne ± un écart-type ; 95% dans un intervalle de ± 2σ ; 99% dans un intervalle de ± 2,6σ et 99,9% dans un intervalle de ±3,3σ (courbe en cloche ou de Laplace-Gauss). De ces propriétés, on tire un risque d'erreur (accepter ou rejeter à tort l'hypothèse d'un usage semblable chez Hugo par rapport à la moyenne de tous les autres). Par exemple, pour les écarts positifs, le risque sera respectivement de : 5% ($z \geq 2$ et $z < 2{,}6$), 1% ($z \geq 2{,}6$ et $z < 3{,}3$), etc.

Pour *homme* chez Hugo, l'écart effectivement constaté entre l'observation et la valeur attendue est 33,4 fois plus important que l'écart-type. On peut donc affirmer, avec un risque d'erreur infinitésimal, que V. Hugo utilise significativement plus ce vocable que la moyenne des autres littérateurs.

Le même calcul est appliqué aux substantifs, aux adjectifs et aux verbes les plus fréquents d'Hugo (tableaux 4 à 6). Pour résumer l'information, les signes ++ et - - indiquent des écarts significatifs (moins de 1 chance d'erreur sur 1000) ; les signes + ou – un risque inférieur à 1% mais supérieur ou égal à 1‰.

Le groupe nominal

Hugo utilise un peu plus de substantifs (et d'adjectifs) que la plupart des auteurs de la bibliothèque. Il partage cette caractéristique avec la majorité des poètes que nous avons étudiés. De ce fait, si l'on calculait les fréquences relatives, en rapportant les effectifs de chaque vocable au total des mots, la plupart des substantifs apparaîtraient sur-employés par Hugo. Il faut donc neutraliser l'influence de cette caractéristique en utilisant comme dénominateur le total des substantifs et non le total des mots (données dans le tableau 3). Ainsi, nous avons la réponse à la question : « quand Hugo utilise un substantif, où vont ses préférences et ses réticences ? ».

Le tableau 4 liste les vingt substantifs les plus utilisés par Hugo avec leurs rangs et leurs fréquences comparés avec le reste de la section littéraire[3].

[3] Pour obtenir ces données de la BEFM : click gauche sur corpus Hugo, sélectionner index, lemma, dans requête : [pos="NOM"], calculer.

Par exemple, dans le corpus Hugo, homme est le substantif le plus employé avec une fréquence de 15,38 occurrences pour 1000 substantifs alors qu'en moyenne chez les autres, il occupe la seconde place (derrière monsieur) avec une fréquence de 10‰. Le z-score (en dernière colonne) indique qu'il n'y guère de chances de se tromper (moins de 1‰) en affirmant que c'est une préférence de Hugo comparé aux autres écrivains.

Le tableau montre que les substantifs les plus fréquents sont souvent les mêmes pour la plupart des auteurs mais que les fréquences d'emploi varient de façon importante. Seuls trois des vingt premiers vocables d'Hugo ne montrent pas d'écart significatif (au seuil de 1%) avec les autres auteurs : *temps, moment* et *coup*. Les deux premiers appartiennent au même champ sémantique de la durée (mais a contrario heure est sous-employé par Hugo).

Tableau 4. Les 20 substantifs les plus utilisés par Hugo comparés à la moyenne des autres littérateurs compris dans la bibliothèque.

Rang Hugo	Rang autres	Vocable	Fréquence Hugo (‰)	Fréquence autres (‰)	Ecart (z)
1	2	homme	15,38	10,00	++
2	1	monsieur	8,98	12,38	--
3	3	Jour	6,96	8,44	--
4	10	Chose	6,56	5,69	++
5	5	Œil	6,25	7,75	--
6	15	Dieu	6,21	5,00	++
7	29	Roi	6,02	2,96	++
8	8	main	5,99	6,57	-
9	21	enfant	5,48	3,94	++
10	11	temps	5,39	5,49	≈
11	7	femme	5,38	7,37	--
12	30	nuit	5,10	3,03	++
13	20	Tête	4,85	4,16	++
14	12	heure	4,62	5,24	--
15	24	âme	4,42	3,54	++
16	342	ombre	4,34	1,26	++
17	19	moment	4,14	4,24	≈
18	17	coup	4,08	4,33	≈
19	52	lettre	4,07	2,11	++
20	25	Ciel	3,86	3,31	++

Il y a des différences anecdotiques comme le sur-emploi de *lettre* (du fait du poids de la correspondance dans ce corpus) ou de *chose* : indice d'un style familier ou « parlé » qu'Hugo utilise spécialement dans son théâtre mais aussi dans sa correspondance et qu'il met dans la bouche de certains personnages de ses romans.

Ombre est le vocable pour lequel l'écart est le plus considérable (Brunet l'avait signalé). Cette singularité n'est pas accidentelle : Hugo sur-emploie aussi *nuit* et *soir* alors qu'il sous-emploie *jour* (ou *lumière*). En descendant dans le détail, on trouvera aussi ces écarts pour : *aurore*, *crépuscule* ou pour certains adjectifs comme *sombre* ou *clair*. Tout cela révèle un goût prononcé d'Hugo pour l'obscurité et, au-delà, pour le mystère.

Beaucoup d'autres singularités font système. Signalons en trois.

Premièrement, le sous-emploi de *monsieur* et de *madame* indique une propension particulière d'Hugo à ne pas faire figurer ces deux mots devant les noms de famille, contre l'usage dominant.

Deuxièmement, et surtout, *madame* qui est au 4e rang chez les autres auteurs, figure seulement au 35e chez Hugo avec une fréquence pratiquement trois fois moindre. Il en est de même pour *femme* (7e rang chez les autres et 11e chez Hugo avec une densité de 20% inférieure) et *fille* (18e chez les autres, 32e chez Hugo, division par deux de la fréquence). Ces caractéristiques sont à mettre en relation avec la forte propension d'Hugo à écrire à propos de l'*homme* et sa réticence à parler du *cœur* (qui passe du 6e rang chez les autres au 23e rang avec une fréquence deux fois moindre) et de l'*amour* (du 9e rang chez les autres auteurs, au 47e chez Hugo avec une fréquence trois fois moindre).

Troisièmement, l'intérêt de Hugo pour *dieu*, l'*âme* le *ciel* (et la *foi*) signale l'importance des thèmes religieux dans son œuvre.

L'examen des adjectifs complète ce panorama (tableau 5).

Tableau 5. Les 20 adjectifs les plus utilisés par Hugo comparés à la moyenne des autres littérateurs compris dans la bibliothèque.

Rang Hugo	Rang autres	Vocable	Fréquence Hugo (‰)	Fréquence autres (‰)	Ecart (z)
1	1	grand	29,83	30,19	≈
2	4	bon	20,85	17,72	++
3	2	petit	17,54	22,25	--
4	7	vieux	13,56	9,26	++
5	5	beau	13,55	15,75	--
6	13	noir	12,04	6,63	++
7	3	seul	11,24	19,04	--
8	9	cher	10,19	8,50	++
9	15	plein	9,10	6,61	++
10	6	jeune	8,71	14,69	--
11	43	sombre	8,09	2,32	++
12	27	profond	7,70	3,48	++
13	16	doux	7,31	6,49	+

14	11	pauvre	7,19	7,22	≈
15	10	vrai	7,07	7,82	-
16	19	haut	6,92	4,53	++
17	18	blanc	6,35	5,41	++
18	44	humain	6,18	2,47	++
19	38	charmant	6,06	2,70	++
20	12	heureux	5,43	7,34	--

Hugo ne se distingue pas de la moyenne des autres auteurs pour l'usage de deux adjectifs qui figurent habituellement dans les premiers rangs : *grand* et *pauvre*. En revanche, il fait preuve d'une répugnance manifeste envers leurs antonymes (*petit* et *riche*). Cette asymétrie se retrouve avec ***haut*** (valorisé) / *bas* (sous-utilisé), ***vieux*** / *jeune*, ***beau*** / *laid*, ***vrai*** / *faux*. Parfois, les deux termes du couple sont surutilisés (*noir* et *blanc*) ou sous-employés ensemble (*heureux* / *malheureux*) marquant ainsi des attirances ou des réticences.

L'écart le plus grand est observé avec *sombre* (3,5 fois plus employé par Hugo que par la moyenne des autres). Relié à la préférence de l'auteur pour l'*ombre* et la *nuit* déjà signalée, cela forme l'un des thèmes singuliers de cet auteur.

Le groupe verbal

Comme pour les substantifs, la faible propension d'Hugo pour les verbes est neutralisée en calculant les fréquences à l'aide du total des verbes. Le tableau 6 donne les résultats pour les verbes les plus utilisés par rapport au reste de la section littéraire de la BEFM.

Tableau 6. Les verbes les plus utilisés par Hugo comparés à la moyenne des autres littérateurs compris dans la bibliothèque.

Rang Hugo	Rang autres	Vocable	Fréquence Hugo	Fréquence autres	Ecart (z)
1	1	être	183,60	138,57	++
2	2	avoir	112,37	108,74	++
3	3	faire	38,32	36,05	++
4	4	dire	29,76	29,78	≈
5	5	voir	19,65	20,61	-
6	7	aller	16,22	16,74	-
7	6	pouvoir	14,16	20,66	--
8	10	venir	12,48	11,44	++
9	8	vouloir	11,02	14,41	--
10	9	savoir	10,99	13,42	--
11	11	prendre	8,33	9,20	--
12	19	mettre	7,24	6,25	++

13	13	falloir	6,84	8,25	--
14	21	passer	6,43	5,58	++
15	16	parler	6,42	7,58	--
16	14	donner	6,40	8,28	--
17	12	croire	5,75	9,08	--
18	25	regarder	5,56	4,72	++
19	17	aimer	5,44	7,61	--
20	22	entendre	5,13	5,11	≈
(…)					
23	12	devoir	4,73	9,16	--

Le rang indique qu'Hugo partage les cinq premiers verbes avec les autres littérateurs. En fait, dès qu'un corpus atteint une grande dimension et une certaine diversité, la même hiérarchie se retrouve.

La forte utilisation de *être* et *avoir* s'explique avant tout par sa préférence pour les récits au passé (spécialement passé composé et plus-que-parfait). Le léger sous-emploi de *voir* est à mettre en relation avec celui du substantif *œil* et le suremploi de *regarder* avec celui de *regard* (plus bas dans les listes). Autrement dit, Hugo ne se contente pas de *voir*, il *observe*.

La principale caractéristique de son système verbal réside dans le sous-emploi très important de cinq verbes particuliers : *pouvoir*, *vouloir*, *savoir*, *falloir*, *devoir*. Pour ce dernier, Hugo l'emploie deux fois moins que la moyenne des autres. En premier lieu, il manifeste une réticence exceptionnelle pour la modalisation évoquée plus haut. Les cinq modalisations sont toutes sous-employées : la possibilité (*pouvoir faire*), la volonté (*vouloir faire*), la compétence (*savoir faire*), la nécessité (*falloir faire*) ou l'obligation (*devoir faire*). Le procédé lui répugne manifestement, mais ce sont aussi les idées associées qui lui déplaisent probablement comme le montre par exemple, le net sous-emploi des substantifs *devoir* ou *volonté*.

Conclusions

Cette rapide discussion montre qu'il ne faut pas considérer les vocables isolément mais reconstituer les combinaisons (syntagmes) dans lesquelles ils entrent et la famille (ou "paradigme") à laquelle ils appartiennent. Les concordances sont un outil précieux pour cela. D'autres ressources permettront d'aller encore plus loin. Ainsi, pourra-t-on prolonger l'esquisse qui vient d'être présentée en reconstituant les champs sémantiques qui structurent le lexique d'un auteur ou celui de la langue.

Ces aperçus sur Hugo ont peut-être paru banaux ou évidents à certains lecteurs familiers de cet auteur. Certes, certains ont pu être énoncés par des critiques ou des universitaires. Mais

le statut de ces caractéristiques change de nature : elles sont maintenant vérifiables et reproductibles au lieu d'être fondées sur une lecture érudite ou sur des intuitions appuyées sur des recensements partiels et bien peu sûrs.

Ajoutons que cette communication ne présente qu'une petite partie des utilisations possibles de la bibliothèque électronique. Le corpus Hugo a déjà été utilisé, notamment pour l'étude des genres littéraires (Labbé & Labbé 2009), l'attribution d'auteur (Labbé 2014). On peut aussi examiner la répartition des mots (et des thèmes), le sens spécifique que donne l'auteur à ses mots favoris, les principales combinaisons de mots, les longueurs et les structures de phrases, approfondir les thèmes, détecter les ruptures et les continuités dans une oeuvre, mesurer la richesse de son vocabulaire et ses singularités. En définitive, il s'agit d'une véritable stylométrie (Savoy 2020) qui contient la lexicométrie classique mais la dépasse singulièrement et dont la plus grande partie reste à inventer.

On l'aura compris, le but de la BEFM est de fournir aux chercheurs des données utiles à propos de la langue, des œuvres ou des auteurs. Elle ne pourra pas se substituer à eux pour la plupart des calculs et des analyses et, naturellement, il leur appartiendra de commenter et de conclure.

Nous espérons parvenir, dans de brefs délais, à mettre en ligne l'ensemble de la Bibliothèque Electronique du Français Moderne telle qu'elle a été constituée en quarante années de recherche. Peut-être, pourra-t-elle apporter une aide aux chercheurs et aux enseignants en sciences humaines et sociales ? Nous sollicitons par avance, la bienveillance des futurs usagers.

Remerciements et crédit.

Nous remercions les organisateurs de cette semaine DATA-SHS et les responsables de la PUD - Grenoble-Alpes.

Une partie des textes utilisés pour constituer le corpus Hugo ont été déchargés sur ABU (association des bibliophiles universels) puis sur Wikisource dont nous remercions les contributeurs.

Frédéric Urien est stagiaire dans l'équipe Sigma (LIG-UGA), chargé de mettre en ligne la BEFM.

Les services informatiques du Laboratoire d'Informatique de Grenoble (Université de Grenoble-Alpes) nous assistent dans la mise en ligne de la Bibliothèque électronique du Français Moderne.

Edward Arnold, Guy Bensimon, Jean-Guy Bergeron, Mathieu Brugidou, Pierre Hubert, Nelly & Jean Leselbaum, Xuan Luong, Thomas Merriam, Denis Monière, Gaétan Péaquin, André Pibarot, Jacques Picard, Mathieu Ruhlman et Jacques Savoy ont collaboré à la mise au point des outils de lexicométrie.

Références

Les *Œuvres de Victor Hugo* (Gallimard, La Pléiade).

Benveniste Emile (1965), "Structure des relations d'auxiliarité", *Problèmes de linguistique générale 2*. Paris, Gallimard (Tel), 1980, p. 177-193.

Benveniste Emile (1970), "L'appareil formel de l'énonciation", *Problèmes de linguistique générale 2*. Paris, Gallimard (Tel), 1980, p. 78-88.

Brunet Etienne (1988a). *Le vocabulaire de Victor Hugo*. Paris : Champion-Slatkine, 1988.

Brunet Etienne (1988b). Hugocentric Tendencies or Can One Approach Hugo Counting Words. *Literary and Linguistic Computing*. 1988, 2, p. 79-9.

Brunet Etienne (1988c). La structure lexicale dans l'oeuvre de Hugo. In Labbé Dominique, Philippe Thoiron et Serant Daniel (dir.). *Etudes sur la richesse et la structure lexicales*. Paris : Champion-Slatkine, p. 24-42.

Heiden Serge, Magué Jean-Philippe, Pincemin Bénédicte (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In Sergio Bolasco, Isabella Chiari, Luca Giuliano (Ed.), Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010) (Vol. 2, p. 1021-1032). Edizioni Universitarie di Lettere Economia Diritto, Roma.

Labbé Cyril & Labbé Dominique (2009). Existe-t-il un genre épistolaire ? Hugo, Flaubert et Maupassant. *In* Banks David. *Le texte épistolaire du XVIIe siècle à nos jours*. Paris : L'Harmattan, 2013, p. 53-85.

Labbé Cyril & Labbé Dominique (2010). La modalité verbale en français contemporain. Les hommes politiques et les autres. In Banks David (dir.). *La modalité, le mode et le texte spécialisé*. Paris : L'Harmattan, 2013, p. 33-61.

Labbé Dominique (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble : Cahiers du CERAT.

Labbé Dominique (2014). Identification de l'auteur d'un texte (Hugo, Lamartine, Musset et Vigny). Conférence invitée au séminaire L'œuvre et son auteur : problèmes d'attribution. Lille : Université de Lille-Nord de la France, 21 mai 2014.

Mandelbrot Benoît (1957). Étude de la loi d'Estoup et de Zipf. Fréquences des mots dans le discours. Apostel Léo et al. *Logique, langage et théorie de l'information*. Paris, PUF, p. 22-53.

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

Muller Charles (1963). Le mot, unité de texte et unité de lexique en statistique lexicologique. *Langue française et linguistique quantitative*. Genève-Paris: Slatkine-Champion, 1979, p 125-143.

Savoy Jacques (2020). *Machine Learning Methods for Stylometry*. Cham : Springer.

Zipf George K. (1935). *La psychobiologie du langage*. Paris : CEPL, 1974.

Sitographie

Les requêtes CQL sur Frantext
https://wiki.frantext.fr/bin/view/Main/Manuel%20d%27utilisation/Expressions/Les%20expressions%20CQL/

Exemples de requêtes CQL (Bibliothèque du français médiéval)

http://bfm.ens-lyon.fr/IMG/pdf/quickref_cql_bfm.pdf

Textométrie http://textometrie.ens-lyon.fr

Text Encoding Initiative https://tei-c.org/

Annexes

1. Le corpus Victor Hugo (1802 - 1885)
(novembre 2020)

	Dates	Longueur (mots)	Vocabulaire (nombre de vocables)
Correspondance (1849-1870)		**292 168**	**9 994**
Poésie*			
Contemplations	1830-1852	91 890	5 935
Châtiments (les)	1853	55 451	5 888
Légende des siècles (la)	1859-1883	215 760	10 278
Total poésie		**363 101**	**12 849**
Roman			
Notre Dame de Paris	1831	185 482	10 746
Misérables (les)	1862	564 294	17 371
Travailleurs de la mer (les)	1866	141 138	9 781
Homme qui rit (l')	1869	199 994	12 380
1793	1874	123 752	8 416
Total roman		**1 214 660**	**26 805**
Théâtre			
Cromwell	1827	81 875	6 858
Hernani	1830	19 578	2 304
Marion Delorme	1831	19 665	2 304
Le Roi s'amuse	1832	15 987	1 957
Marie Tudor	1833	22 130	1 889
Lucrèce Borgia	1833	19 778	2 186
Ruy Blas	1838	24 198	2 984
Torquemada	1869	20 037	2 729
Total Théâtre		**223 248**	**9 622**
Total		**2 093 177**	**32 665**

* Données provisoires. Ce corpus est en cours de correction. De plus, les *Odes et ballades* et *l'Art d'être grand-père* seront intégrés dans quelques jours.

2. Réglage des informations visibles concernant les textes actuellement disponibles dans la BEFM en ligne.