L'EXPLOITATION DES DONNÉES DU TLF

Etienne BRUNET

Depuis plus de vingt ans qu'on amasse à Nancy des données pour le *Trésor de la langue française ,* le précieux coffre qui les renferme est plein à ras bords. Celui qui soulèverait le couvercle y dénombrerait plus de 150 millions de pièces, je veux dire de mots, et quelque 3000 textes complets de la littérature française, du dix-septième siècle à nos jours. Certes ce trésor n'a pas été enfoui dans une cave souterraine, et la communauté scientifique connaissait son existence. En dehors des rédacteurs du dictionnaire, certains chercheurs ont pu puiser à la mine de Nancy - et nous avons été parmi les premiers bénéficiaires. Mais les prestations de services extérieurs ne pouvaient être prioritaires tant que l'achèvement du TU n'était pas acquis. Et les demandes étrangères à la rédaction étaient plus facilement satisfaites quand elles coïncidaient avec les traitements de routine qui pour telle ou telle lettre de l'alphabet exigeaient la lecture linéaire de tous les textes enregistrés. Ces opérations séquentielles, longues et coûteuses, n'étaient renouvelées qu'à de longs intervalles et le délai entre la demande du chercheur et la réponse obtenue était trop long pour qu'un véritable dialogue pût s'instaurer et que les questions pussent être affinées et soumises derechef à la machine.

Ce dialogue est maintenant possible. Les données de Nancy ont été intégrées dans une base de données qui est actuellement au monde non seulement la plus vaste, mais aussi la plus rapide et la moins coûteuse parmi celles qu'on peut. consulter dans le domaine linguistique et littéraire. Il ne nous appartient pas d'en détailler les principes et le mode opératoire et nous renvoyons le lecteur au créateur de cette base, J. DENDIEN [1]. Bornons-nous à dire que ce puissant logiciel porte le nom de *STELLA,* et qu'il est accessible par le réseau TRANSPAC de tout point de l'hexagone, et même à l'extérieur des frontières. Nul besoin de demande

1 Un système de gestion de base de données textuelles, *Méthodes quantitatives et informatiques dans l'étude des textes,* Slatkine-Champion, Genève-Paris, 1986, t.1, pp.285-293.

écrite, et plus de ces retards propres au batch et au différé et de ces malentendus propres aux prestations de service. Le chercheur , au bout de sa ligne télématique, peut pêcher sans attendre n'importe lequel des 150 millions de mots du grand corpus, s'intéresser à une expression, à un écrivain, à un genre littéraire, à une époque, ou à tel ensemble de textes qu'il précise et corrige à sa guise. Et la réponse lui est fournie immédiatement (en quelques secondes), sous la forme qu'il a choisie et qu'il peut modifier en cours de route: des fréquences, des références, ou des contextes de la longueur qu'on veut et dans l'ordre qu'on désire. Le chercheur qui dispose de possibilités de stockage peut enregistrer les informations reçues et les exploiter par la suite ., en les soumettant à des traitements locaux. Mais le plus souvent les ressources du langage d'interrogation sont si larges qu'il n'est pas nécessaire de recourir à des opérations ultérieures de complément. Pour donner une idée des possibilités de la base STELLA, et sans sortir du domaine de la lexicométrie qui nous est habituel, nous proposerons le tableau ci-dessous (figure l),qui a été acquis sans grande dépense et qui établit les effectifs pour l'étude comparée du bestiaire de quelques écrivains. La liste des auteurs choisis et celle des mots retenus relève de la liberté du chercheur, qui circonscrit comme il l'entend son corpus et son champ de recherche, dans des limites très larges (le corpus exploré dans le tableau 1 contient plus de 15 millions de mots). On trouvera ailleurs les conclusions qu'un tel tableau, convenablement pondéré, peut inspirer. L'essentiel est de montrer qu'avec STELLA une porte est largement ouverte pour des recherches ou des vérifications jusque là impossibles, que le point de vue soit sémantique, morphologique ou syntaxique.

Cette puissance, cette instantanéité, et cette souplesse de STELLA rendent un peu désuets les logiciels qui ont été appliqués aux données du Trésor et qu'on a vu naître dans le passé, à Nancy et ailleurs. Dans une discipline qui évolue très vite et qui propose chaque jour des machines plus performantes et des méthodes plus rapides, les réalisations qui ont un passé n'ont plus d'avenir et il n'est pas difficile de prévoir que les menaces pèsent sur STELLA au moment même de sa naissance. Est-ce que la télématique doit rester longtemps le canal obligé des bases de données? La technologie du laser permet d'envisager une autre distribution des grandes masses d'information, et cela permettrait d'échapper aux contraintes des liaisons à distance lesquelles sont parfois perturbées, parfois indisponibles et toujours coûteuses. Il est fort possible que dans un proche avenir les données du Trésor soient enregistrées sur un support à lecture optique (un CD-ROM) et que leur exploitation soit

assurée sur les sites locaux, à l'aide d'un simple micro-ordinateur. Aucun obstacle technique n'interdit de s'engager dans cette voie, pas même l'étendue du corpus, puisque deux disques de 12 cm de diamètre pourraient emmagasiner la totalité des textes de la base. Les freins qui s'opposeront un temps à cette divulgation sont d'ordre juridique et commercial. A supposer que ces difficultés soient résolues, on verra cependant coexister les deux systèmes d'interrogation, la voie télématique restant le recours des chercheurs ou des équipes qui ne disposeront pas d'installations propres. Et de la même façon la réalisation présente de STELLA n'abolit pas tout-à-fait les réalisations antérieures, qui se maintiendront quelque temps encore, parce qu'elle répondent à des besoins spécifiques que STELLA ne peut satisfaire commodément.

C'est le cas en particulier des recherches lexicométriques portant sur tous les mots d'un corpus. STELLA demande en principe au chercheur de préciser les mots auxquels il s'intéresse. Certes l'emploi d'un critère très large de filtrage (l'option de la commande mots) permet de retenir tout le vocabulaire du corpus courant. Mais la sortie des résultats sera alors fort volumineuse et l'on peut craindre le coût et la précarité de leur transmission. En son état actuel, STELLA ne s'engage pas plus avant dans l'exploitation statistique et, de toutes façons, dans ce domaine, le relais est passé au chercheur. Il peut dès lors paraître avantageux de remonter à la source où STELLA a puisé ses données, c'est à dire aux fichiers-répertoires (ou F.R.) qui contiennent les index de tous les textes anciennement dépouillés. Comme ces FR se trouvent disponibles sur bandes magnétiques, la prestation de service se réduit à une simple copie. La structure de ces fichiers n'est pas un modèle de simplicité, mais on en vient à bout sans difficulté en assembleur, ou aussi bien dans des langages évolués comme le Cobol ou le PLI. Les enregistrements sont de longueur variable, mais pour une même espèce d'information (étendue des pages, références, sous-références, fréquences, localisations, graphies) la longueur est toujours la même, et toujours différente de celles des autres espèces - et cela suffit à établir la nature de l'enregistrement. Voici comment s'établit la chaîne des traitements institués à Nice, dans l'URL 9 Etude statistique du Trésor littéraire :

	Colette	Balzac	Chatea	Claude	Giraud	Hugo	Maupa	Pergau	Proust	Renard	Zola	TOTAL	
abeille	11	10	31	68	34	74	1	0	5	3	13	250	1513
aigle	0	68	104	44	24	303	4	0	15	11	33	606	1930
âne	3	34	21	73	24	80	15	0	8	34	63	355	1724
animal	25	142	127	201	201	99	114	8	68	42	118	1145	12839
araignée	12	16	7	30	16	96	11	0	4	36	39	267	904
boeuf	3	39	43	82	32	123	34	6	30	92	90	574	2438
canard	0	41	11	10	22	12	22	0	10	35	39	202	807
cerf	0	18	25	19	34	20	11	1	4	5	24	161	558
chat	198	125	31	29	95	138	80	7	16	81	263	1063	3471
cheval	50	485	280	241	235	533	252	2	115	107	656	2956	14521
chèvre	5	25	30	26	14	124	11	0	6	11	50	302	1256
chien	97	185	103	180	318	268	289	42	41	155	384	2062	8180
chouette	3	4	5	7	24	19	11	3	6	10	13	105	335
cochon	7	22	3	42	5	10	48	0	5	69	254	465	1602
coq	1	25	11	37	29	40	30	6	11	46	46	282	1155
cygne	1	18	51	22	40	85	6	0	20	9	18	270	995
dragon	2	13	35	34	14	102	12	0	8	2	19	241	1024
écureuil	1	3	10	4	23	4	5	3	2	6	2	63	306
éléphant	1	6	17	16	29	66	7	0	0	7	3	152	926
fourmi	2	6	8	14	29	25	14	0	0	14	25	137	728
gibier	2	20	1	13	34	5	20	1	4	11	20	131	699
insecte	5	38	47	27	46	30	12	10	38	10	52	315	2496
lapin	11	14	10	31	14	8	65	4	7	30	101	295	1214
lièvre	8	6	16	14	27	10	20	35	1	36	29	202	824
lion	0	67	104	92	31	276	10	0	16	19	33	648	2580
loup	11	74	53	32	30	165	51	1	20	19	115	571	2301
moineau	1	20	6	5	24	43	3	2	2	37	53	196	447
mouche	21	61	19	35	34	124	35	3	12	57	84	485	1992
mouton	8	49	26	53	9	49	29	0	5	42	71	341	1768
oie	5	8	9	7	21	26	10	0	5	48	102	241	704
oiseau	45	107	279	214	358	652	126	42	105	119	158	2205	9928
ours	2	53	69	36	11	98	5	0	4	16	31	325	1259
papillon	13	15	14	46	14	77	3	0	21	15	28	246	1141
perroque	1	13	5	9	25	6	9	0	3	11	1	83	472
pigeon	2	25	13	21	20	15	20	0	22	23	47	208	909
poisson	7	48	47	154	65	61	71	1	34	56	103	647	3414
poule	14	37	17	26	31	39	63	15	16	64	116	438	1442
poulet	14	27	6	11	13	7	43	0	15	23	69	228	792
rat	8	51	9	29	32	75	8	2	9	19	30	272	1270
renard	6	10	7	16	23	23	23	30	0	10	8	156	830
rossignol	2	19	43	22	30	21	6	0	3	5	2	153	810
serpent	11	44	113	40	43	79	35	4	23	21	28	441	1961
singe	4	46	16	14	34	39	19	0	9	17	24	222	1360
souris	14	19	17	29	34	99	13	1	12	19	31	288	861
taureau	0	21	39	23	21	43	2	0	3	28	23	203	1425
tigre	3	68	46	23	27	134	6	0	4	3	2	316	860
truite	0	2	2	5	37	4	8	0	6	2	10	76	224
vache	12	22	37	57	19	42	54	2	16	83	162	506	2084
veau	9	21	9	22	21	8	11	0	5	40	83	229	808
TOT2	661	2290	2032	2285	2370	4479	1757	231	794	1658	3768	22325	1E+05

FIGURE 1. Le BESTIAIRE comparé de quelques écrivains
Effectifs obtenus par la base de données STELLA

a - La première opération consiste à tirer de ce fichier original une copie de travail dans le format standard.

b - Les données appartiennent à une édition de référence qui n'est pas nécessairement celle qu'on souhaiterait. Or un index ou une concordance ne sont utiles que si les références renvoient à une édition disponible et irréprochable. Et c'est la collection de la Pléiade qui réunit le plus souvent ces deux qualités. On est ainsi amené à reconstituer le texte, à partir de l'index fourni .Et sur le listing précisément jalonné qu'on obtient, il reste à retrouver les ruptures de page de l'édition de la Pléiade. La moitié des Rougon-Macquart (soit 3000 pages) et la majorité des textes de Hugo (soit 5000 pages) ont été ainsi reconstitués. Cette phase du traitement comporte un tri, et une édition de travail, puis un repérage manuel et la transmission, après contrôle, de ces repères à la machine.

c - On reconstitue alors, sur bande, un nouvel index, qui contient, pour chaque forme de chaque texte, la transcription en alphabet pauvre et en alphabet riche, la fréquence et la suite chronologique des références de page(dans la nouvelle édition) et des zones dans la page (codées de a à g).

d - Reste à restituer la notion de texte. Les textes-machines ne sont qu'une division arbitraire imposée par d'anciennes contraintes de la technologie (tranches de 100 000 mots maximum). Il convient de coller les morceaux qui appartiennent au même ensemble. Et c'est l'occasion d'un tri (qui prend en compte d'abord l'ordre alphabétique, puis, en critère secondaire, le rang chronologique du "morceau" de texte), suivi d'un programme de tassement.

e - Vient alors la phase cruciale de la lemmatisation. Le fichier issu du tassement est confronté à un modèle, fourni par l'URL 1 de Nancy sous le nom de CODGB. Ce modèle propose un lemme de rattachement à toutes les graphies rencontrées. Cette information est transmise au corpus étudié, ainsi qu'un code grammatical également présent dans le modèle. Il ne faut pas se cacher que cette lemmatisation automatique est assez expéditive et que pour dissoudre toutes les homographies il faudrait un recours (nécessairement "manuel") au texte. La phase de lemmatisation est suivie d'un tri, qui ordonne les enregistrements selon l'ordre alphabétique du lemme, puis de la forme, puis selon l'ordre chronologique.

f - On constitue à ce moment le dictionnaire des fréquences pour le corpus étudié . Et cette opération comporte plusieurs traitements:

- Le premier reprend le fichier précédent en abandonnant les références et les formes, et en ne gardant que les lemmes , les fréquences et la distinction des textes. Mais cette distinction des textes ne tient plus à un code particulier inscrit dans des enregistrements juxtaposés, elle apparaît sous forme d'un tableau de fréquences à l'intérieur d'un enregistrement cumulatif qui regroupe toutes les informations relatives à un même vocable. En même temps on constitue les effectifs des vocables et des occurrences rencontrés dans chaque sous-ensemble et dans le corpus (en écartant, grâce au codage grammatical, les noms propres et les mots étrangers).

- Ces effectifs servent à établir les probabilités qui permettent l'exploitation statistique du corpus. Le dictionnaire des fréquences est alors complété par des indications comparatives qui convertissent en écarts réduits les variations constatées pour un même mot dans la suite des textes(à partir d'un certain seuil de fréquence). Divers tests statistiques sont effectués qui mesurent, pour chaque mot , la corrélation chronologique, la répartition selon le genre, la dispersion.

- Enfin cette comparaison interne est doublée d'une confrontation externe. En prenant appui sur le dictionnaire des fréquences de l'ensemble du corpus XIX-XX e siècle de l'INaLF, d'autres écarts réduits sont calculés qui permettent de préciser la spécificité du corpus étudié. La comparaison se fait avec un ou plusieurs corpus de référence, les contraintes de genre ou d'époque permettant de rapprocher les semblables pour observer leurs différences.

- Ainsi achevé sur bande magnétique, le dictionnaire des fréquences est enfin livré à l'impression, dès lors qu'on été établis , par programme, les codes typographiques destinés à la photocomposeuse ou à l'imprimante à laser.

g - Le dictionnaire des fréquences sert aussi à la fabrication de l'index définitif, qui cumule toutes les informations statistiques issues de la phase f et toutes les références établies dans la phase e. De ces deux fichiers, consultés conjointement, un programme complexe d'édition extrait un index synoptique qui pour un même vocable délivre dans l'ordre alphabétique les diverses formes rencontrées. Tous les sous-ensembles du corpus étudié y sont mis en parallèle, chacun disposant d'une colonne qui contient tout à la fois les fréquences, les

écarts et les références. Selon les cas et l'importance du corpus, un tel index est imprimé sur papier ou bromure (après codage typographique) ou préparé sur bande en vue d'un transfert sur microfiche.

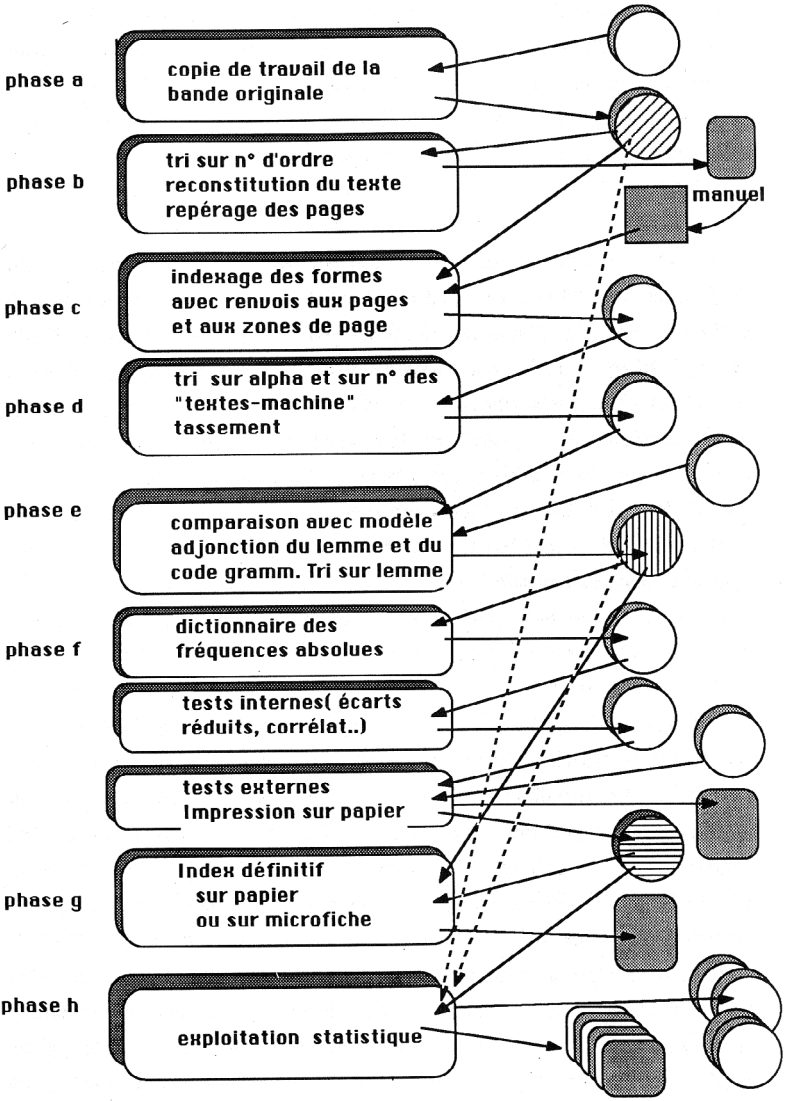
h - Quand ces outils de base ont été constitués, l'exploitation statistique emprunte des chemins fort variés qui font appel à des programmes spécifiques et dont la multiplicité décourage l'énumération. Ces programmes, par exemple, constituent le tableau de distribution des classes de fréquence, relèvent les hapax, établissent l'index hiérarchique, évaluent l'étendue du vocabulaire, étudient la répartition des catégories grammaticales, des mots de relation, des signes de ponctuation, des suffixes, des préfixes, des champs sémantiques, et de tout objet d'étude qu'on prétend isoler. Ils permettent les comparaisons interne et externe, la mesure de l'évolution, et de la spécificité. Et ils fournissent les données numériques aux logiciels d'analyse factorielle.

En résumé le processus peut être visualisé dans le schéma reproduit ci-dessous (figure 2).

Précisons que ce processus ne concerne que les données issues du TLF, et plus précisément des fichiers de base dits fichiers-répertoires. D'autres traitements ont été prévus à l'URL 9 pour les textes qui ont une source différente et auxquels s'appliquent des règles différentes de lemmatisation ou de codage grammatical (notamment les textes anglais et latins). Ajoutons aussi que cette chaîne de traitements utilise la télématique, principalement le puissant serveur de Montpellier, et qu'elle concerne les grands corpus dont le laboratoire s'est fait une spécialité.

Mais on n'écarte pas pour autant les petits corpus qui, eux, peuvent être traités intégralement sur place, y compris la saisie. Si le texte étudié garde des proportions restreintes (moins de 500 pages), toutes les opérations peuvent être assurées par des micro-ordinateurs locaux, pourvu que la capacité de stockage soit suffisante (il est nécessaire de disposer de disques durs de 20 ou 40 millions de caractères)

Une chaîne de traitements a été mise en place pour ces applications locales, qui aboutissent non seulement à la fabrication d'Index mais aussi de Concordances. Ces traitements ont recours aux langages disponibles en microinformatique, c'est-à-dire au Basic et au Pascal, tandis que les gros traitements précédemment décrits font appel au langage PL/1 (et parfois au Cobol).



FIGUR

E 2. SCHÉMA de traitement des fichiers-réperloires

. Enfin la saisie même des textes peut être assurée sur place, par recours à la lecture optique (les scanners et les logiciels de reconnaissance des formes étant devenus d'un prix abordable).

Faut-il suivre l'expérience d'un laboratoire qui s'est lancé parmi les premiers dans la programmation et qui depuis quinze ans a réalisé des centaines de programmes spécialisés dans le traitement des textes? S'engager dans cette voie, c'est se condamner au mouvement perpétuel. Car il faut suivre la progression des machines, adopter le langage qu'elles veulent entendre, et réécrire parfois la chaîne des traitements, parce que les circonstances imposent du matériel nouveau ou des méthodes inédites. L'adaptation aux changements exige une démarche souple qui conduit à préférer les unités de programmation légères, morcelées, autonomes, aisément contrôlables, plutôt que les lourds "packages" intégrés, à l'image de ceux qu'on a faits à Montréal (logiciel *JEUDEMO)* ou à Oxford (COCOA). Car de telles chaînes de traitements sont contraignantes pour le chercheur et fort dépendantes d'une installation particulière. Et leur créateur doit fournir une maintenance sans faille et une documentation sans erreur.

STELLA ,qui jouit du mode conversationnel, n'a pas la lourdeur des packages précités, et son utilisation est fort conviviale, même pour les non-initiés. Sa puissance permet de donner réponse à la plupart des questions qu'on peut poser décemment à une machine. Mais certains développements y demeurent en suspens, notamment du côté de la statistique lexicale. Tant que ces compléments ne seront pas disponibles, la programmation restera nécessaire.

Etienne BRUNET
URL 9. Institut National de la langue française