LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI
F. BENZÉCRI

Sur la comparaison entre deux corpus analysés, d'abord, séparément

Les cahiers de l'analyse des données, tome 20, n° 1 (1995), p. 67-78

<http://www.numdam.org/item?id=CAD_1995__20_1_67_0>

Les Cahiers de l'Analyse des Données
Vol. XX - 1995 - n°1 - pp. 67-78.

SUR LA COMPARAISON ENTRE DEUX CORPUS ANALYSÉS, D'ABORD, SÉPARÉMENT

[COMPAR. CORPUS]

J.-P. & F. BENZÉCRI

1 Origine de la présente étude

Dans [TEXTES RUSSES], in *CAD*, XIX, n°1, est publiée l'analyse d'un corpus JF de 62 fragments de textes littéraires ou philosophiques en langue russe. Dans [COMPAR. RUSSE], in *CAD*, XX, n°1, paraît une étude approfondie de l'œuvre de M.A. CHOLOKHOV, dans ses rapports avec celle d'un autre auteur, F.D. KRIUKOV, dont on a pu conjecturer qu'il était le véritable auteur d'une partie de ce qui a été publié sous le nom de CHOLOKHOV.

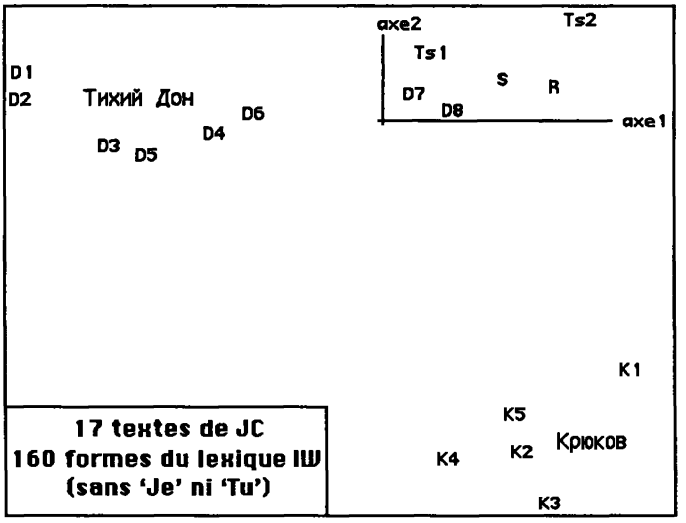
Quant à l'importance du travail d'élaboration linguistique, cette deuxième étude l'emporte de beaucoup sur la première: il s'agit, non de 62 fragments (dont chacun ne compte guère plus de 3k caractères), mais d'un ensemble JC de 17 véritables textes ou parties (dont l'ordre de grandeur est 50k), et dont plusieurs sont considérés avec leur subdivision en une vingtaine de chapitres. De plus, dans [COMPAR. RUSSE], les auteurs ne se sont pas bornés à dénombrer des formes graphiques, ou séquences de caractères: ils ont résolu des homonymies et considéré, quand il y avait lieu, des mots composés.

Malgré ces différences, et en partie précisément à cause de celles-ci, il nous a paru qu'une comparaison entre les deux études servirait, à la fois, pour la connaissance de la langue littéraire russe et le progrès de la méthode statistique; et nous croyons que les résultats obtenus méritent de faire l'objet d'une note.

2 Préparation des données

L'étude [COMPAR. RUSSE] étant la plus approfondie, il s'imposait de partir des données de celle-ci: on a pris pour base un tableau 224×17, croisant un lexique V de 224 formes avec un ensemble JC de 17 œuvres (ou parties).

Ainsi qu'on l'a dit, V fait des distinctions qui ne sont pas comprises dans le traitement informatique de [TEXTES RUSSES], il a donc fallu, non sans regret, effacer ces distinctions, dont voici des exemples: {чем ≠ чём}; {что-



ТО ≠ ЧТО + ТО}; ...: ЧЕМ a été cumulé avec ЧЁМ ; et ЧТО−ТО étant supprimé, les occurences en sont comptées deux fois : d'une part avec ЧТО; de l'autre avec ТО.

De plus, des formes de V, certaines sont absentes du corpus des 62 fragments; d'autres s'y rencontrent moins de 5 fois: ces formes ont été supprimées du lexique retenu pour la comparaison.

D'où un lexique de 160 formes, noté IW, qui se prête à des relevés simultanés sur les deux corpus aboutissant à un tableau : IW × (JF ∪ JC), 160 × (62+17).

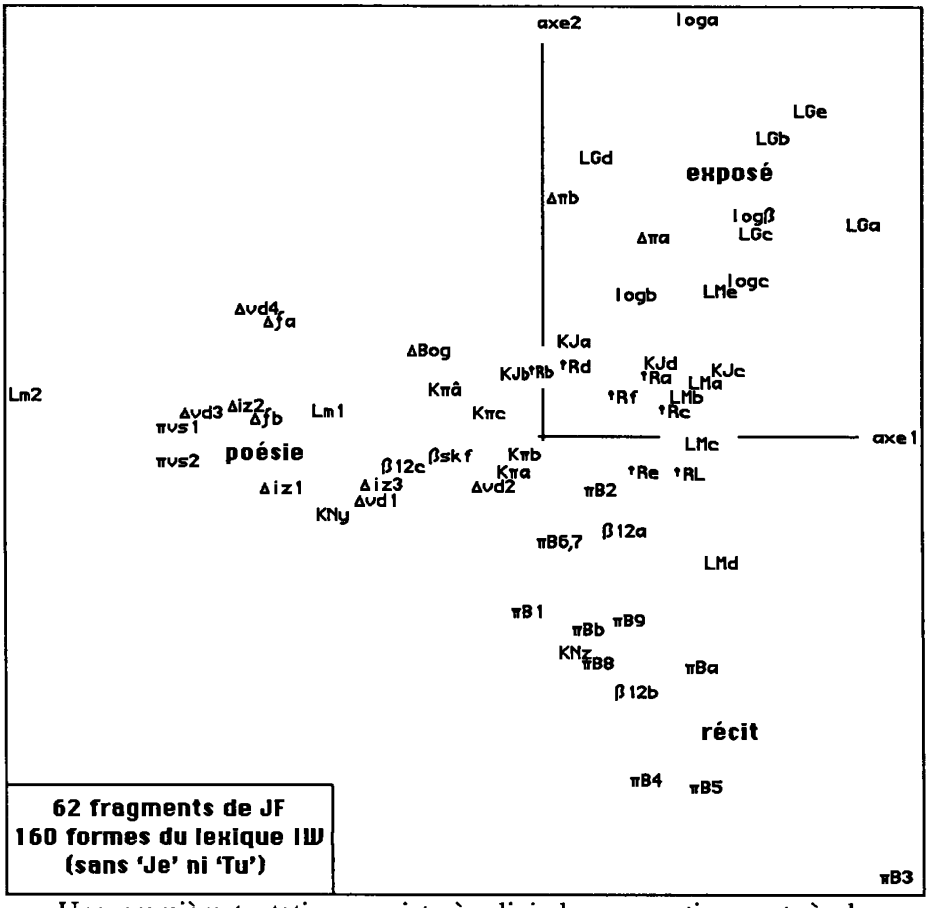
3 Analyses séparées des deux corpus croisés avec le lexique IW

Le lexique IW, n'a pas été choisi d'après le corpus JF des 62 fragments: il a seulement été allégé pour éviter que figurent, dans IW × JF, des lignes trop légères, au profil indéterminé. Avec JC, l'adéquation de IW est, *a priori*, meilleure: mais on a effacé des distinctions intéressantes et éliminé certaines formes (rares, il est vrai, même dans JC).

Il n'est donc pas superflu de noter que les analyses séparées de IW × JF et de IW × JC fournissent, dans le plan (1, 2), comme dans les CAH (non publiées) des résultats analogues à ceux publiés dans les deux articles cités (auxquels nous renvoyons pour l'inventaire détaillé des textes de JF ∪ JC, avec leurs sigles).

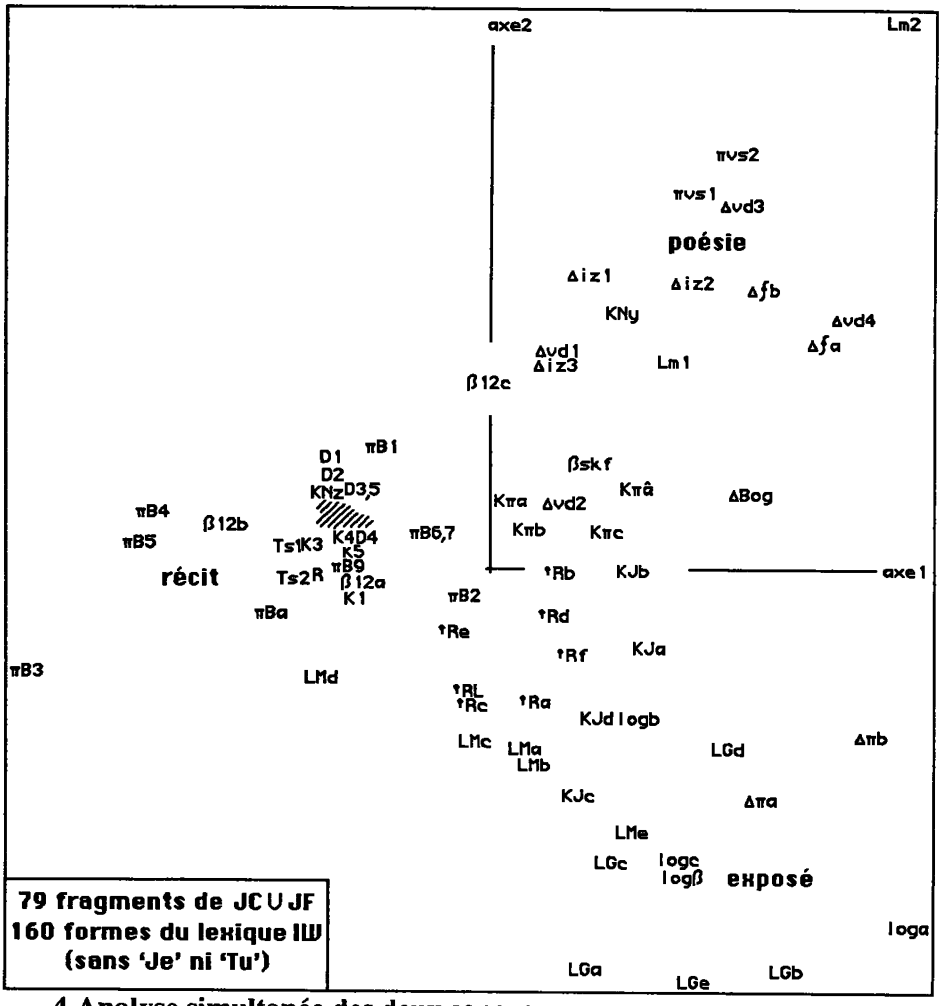
On présume donc que le lexique IW peut, en première approximation, être retenu pour une taxinomie générale des textes en langue littéraire.

Reste la comparaison des deux corpus.



Une première tentative consiste à adjoindre, respectivement, à chacune des analyses IW × JC et IW × JF, l'autre corpus comme ensemble de colonnes supplémentaires. Pour dépouiller les résultats, le plus simple est de considérer le listage des facteurs, en prêtant attention à la qualité de représentation des éléments supplémentaires; en complétant cet examen par une analyse discriminante (laquelle consiste ici, simplement, à déterminer de quel élément principal chaque élément supplémentaire est le plus proche).

Sans entrer dans les détails, nous dirons que, des trois branches {poésie, exposé, récit} reconnues dans l'analyse des 62 fragments de JF, c'est la branche 'récit' qui montre une nette affinité avec le corpus JC des "écrivains du Don"; tandis que les branches {poésie, exposé} n'ont pas d'homologue dans JC. Ce premier résultat, sans surprendre, mérite d'être noté.



4 Analyse simultanée des deux corpus

```
IW × JC ∪ JF : 160 formes × 79 fragments
trace  :   1.453e+0
rang   :     1     2     3     4     5     6     7     8     9    10
lambda :   941   901   682   645   592   501   486   443   412   402 e-4
taux   :   648   620   470   444   407   345   335   305   284   276 e-4
cumul  :   648  1268  1737  2181  2589  2933  3268  3573  3857  4133 e-4
```

Ainsi qu'on l'a souligné, le corpus JF est léger, en comparaison avec JC. Analyser tel quel le tableau global IW × (JF ∪ JC) aboutirait donc à des résultats tendant vers ceux obtenus en mettant, franchement, JF en supplément. Afin de mettre quelque équilibre entre JC et JF, on a multiplié par

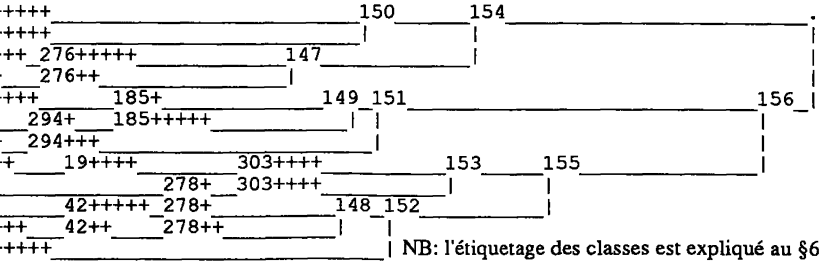
100 les colonnes de ce dernier. Le coefficient 100 ne s'impose pas: il a été choisi pour donner à JF, très dispersé, une masse supérieure à celle de JC, plus concentré. Les résultats dont nous rendons compte ci-après ne dépendent pas de cette valeur précise.

Le plan (1, 2), seul publié, montre les trois branches {poésie, exposé, récit}; le corpus JC s'intègre au récit pour former un agrégat très concentré. La qualité de représentation (dans l'espace des dix premiers axes) est bonne pour l'un et l'autre corpus.

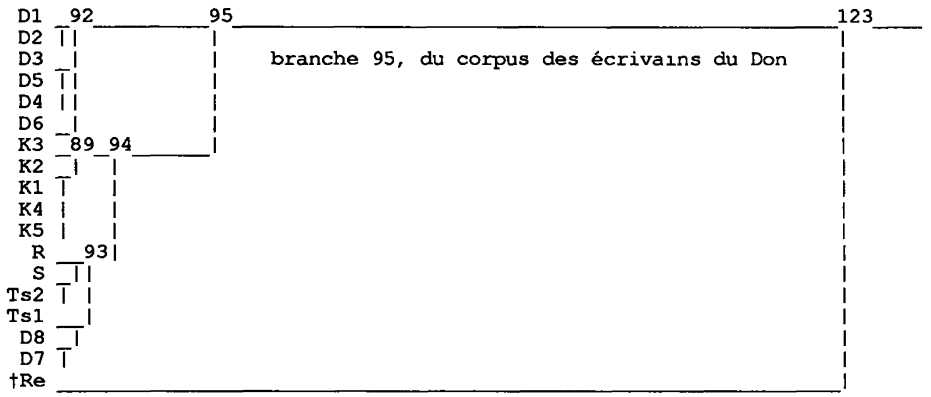
Dans la CAH, une partition en 12 classes, définie par les 11 nœuds supérieurs, montre d'abord la structure connue de JF; JC étant dans la branche 154 qui comprend, principalement les fragments d'un récit de POUCHKINE, πB, "*La demoiselle paysanne*".

Plus précisément, JC est dans la subdivision 145, avec deux fragments de la biographie de ROUBLOV, †R; trois de πB... Si, grâce à un changement d'échelle honrizontale, on rend visible le détail de l'arborescence, il apparaît que les textes de JC s'agrègent entre eux suivant une structure hiérarchique

c	Partition en 12 classes : Sigles des textes de la classe c	
145	D1 D2 D3 D5 D4 D6 K3 K2 K1 K4 K5 R S Ts2	
	Ts1 D8 D7 †Re †Rb πB4 ß12b πB5 πB3 JC avec †Re, etc.	
46	ß12a	
129	KNz πB8	
139	πBb πBa πB9 πB7 †Rc πB1 †RL πB2 πB6	récit de πouchkine
141	Kπc KNy Kπa †Rf	prose de Karamzine
118	Δfa Lm2 Δfb	
138	ΔBog Kπb Δvd1 Δiz2 Δiz3 KJb Δvd2 ßskf Δiz1 Kπâ πvs1 Lm1 Δvd3 πvs2	
	Δvd4	poésie de Δerjavine
142	ß12c LGd	
146	KJa †Ra †Rd logb logß loga LGe LGc	philosophie
74	Δπb	prose de Δerjavine
143	logc LMa KJc KJd LGa LGb Δπa	et
127	LMd LMe LMc LMb	philosophie



NB: l'étiquetage des classes est expliqué au §6



identique à celle trouvée dans [COMPAR. RUSSE]; le niveau d'agrégation étant toutefois très bas, relativement à celui auquel l'ensemble de JC (ici classe 95) s'agrège à un premier fragment de JF, †Re, pour constituer une subdivision 123 de la classe 145.

Pour faire image, on dira que JC, relativement à JF, est concentré en un point; mais en un point dont un grossissement approprié révèle la stucture exacte.

5 Commentaire critique et analyse complémentaire

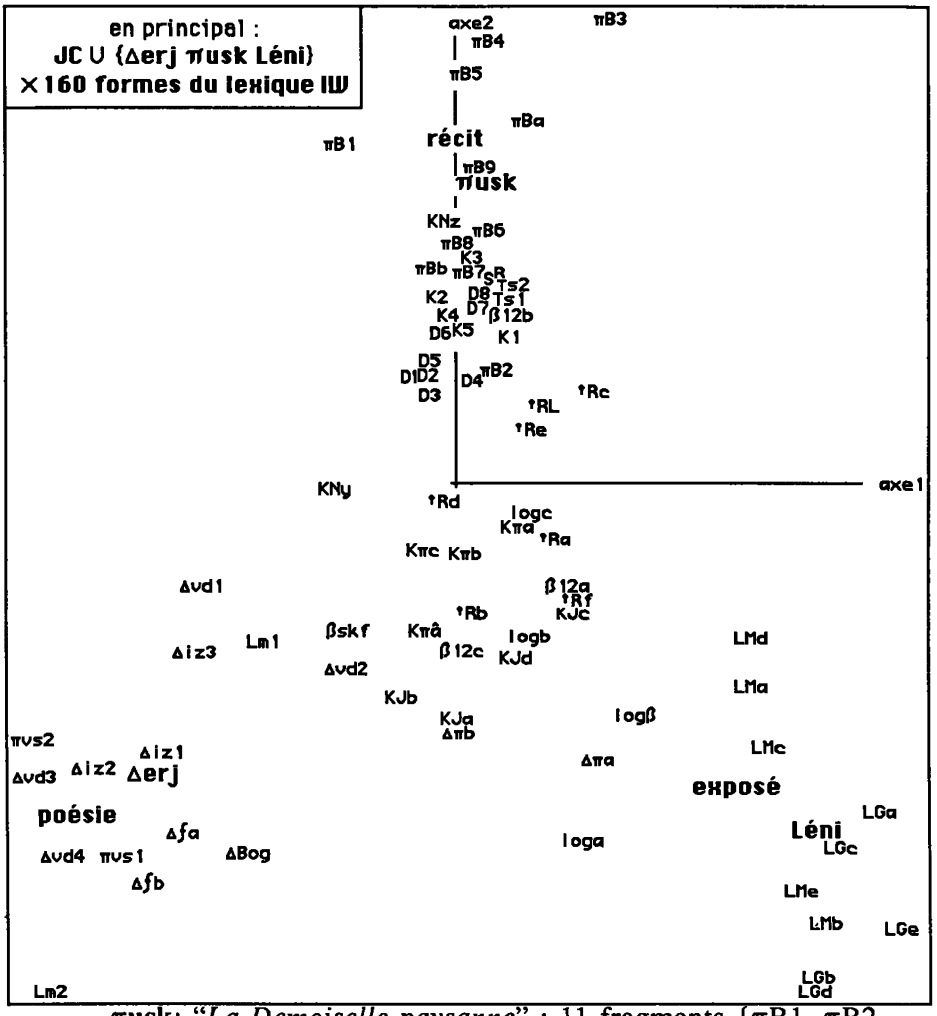
Comment expliquer cette différence d'échelle entre JC et JF? Assurément, par la diversité des genres de ses 62 fragments, JF est, *a priori*, plus étendu que JC dont les 17 textes sont des récits, éventuellement coupés de dialogues (comme le sont aussi des fragments de JF).

Mais il faut aussi prendre garde à la brièveté des fragments de JF, dont le profil lexical est, nécessairement, en butte à des fluctuations d'échantillonnage; tandis qu'au contraire, les longs textes de JC ont des profils où s'estompent non seulement les fluctuations d'échantillonnage mais aussi de véritables variations de style.

Ceci dit, que doit-on craindre? Pour JC, d'avoir perdu le signal en filtrant le bruit; pour JF, d'avoir assourdi celui-là par celui-ci.

Relativement à JC, le §5 de [COMPAR. RUSSE], où sont présentés des graphiques distinguant les chapitres de divers textes, atteste que la structure n'a pas été réduite jusqu'à perdre son sens: la gradation au sein des textes de CHOLOKHOV est confirmée, ainsi que l'opposition entre CHOLOKHOV et KRIUKOV.

Quant à JF, il ne comprend que des fragments; mais dont certains peuvent être cumulés pour montrer les profils d'œuvres homogènes. On a donc repris le tableau IW × (JF ∪ JC), tel qu'il a été analysé au §4 en multipliant par 100 le bloc JC, et créé, par cumul, trois colonnes:



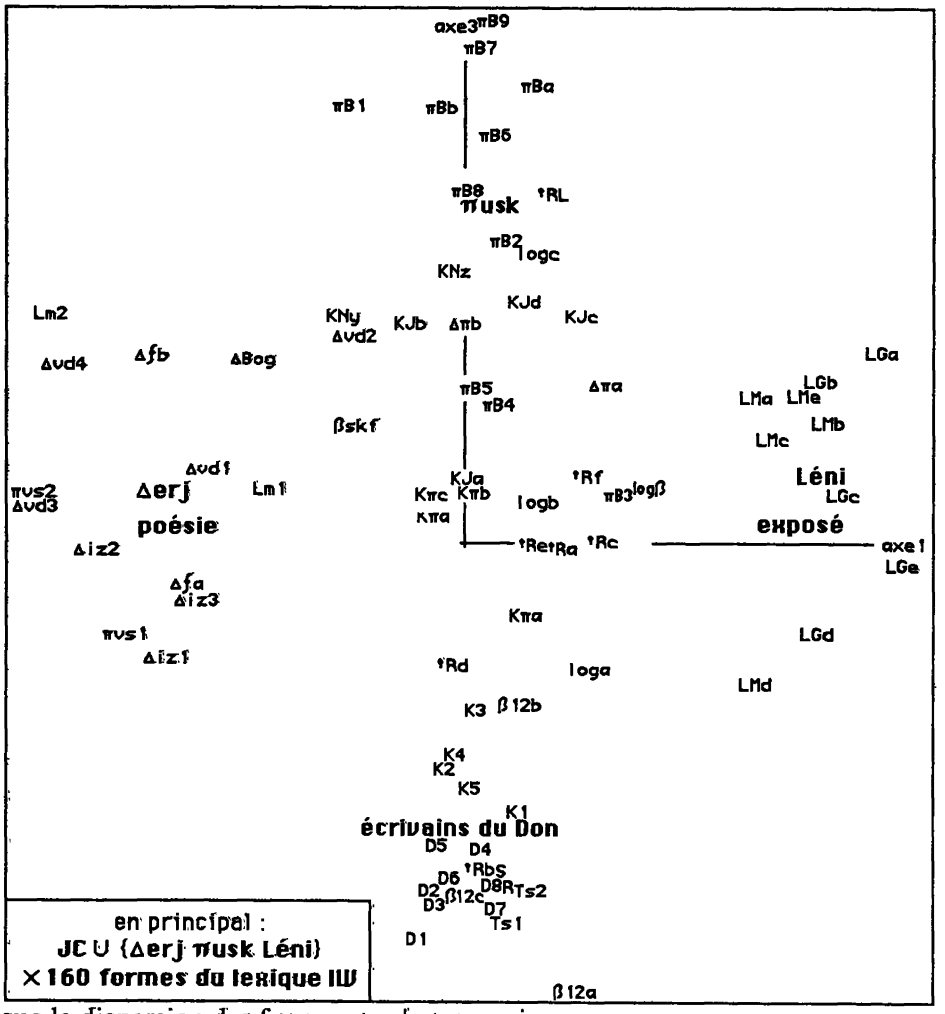
πusk: "*La Demoiselle paysanne*" : 11 fragments {πB1, πB2,... , πBa, πBb};

Δerj: 12 fragments: quatre grandes odes de DERJAVINE, avec l'*Évocation*, ode de POUCHKINE, {πvs1, πvs2};

Léni: 10 fragments de LÉNINE: cinq de "*L'État et la Révolution*", LG; cinq de "*Matérialisme et Empiriocriticisme*", LM.

On effectue une analyse complémentaire; avec, en principal, d'une part JC, d'autre part {Δerj, Léni, πusk}; JF étant en supplément.

Apparaît d'abord l'image du plan (1, 2), avec les trois branches {poésie, exposé, récit}; sous-tendue par {Δerj, Léni, πusk + JC}; et confirmée, alors



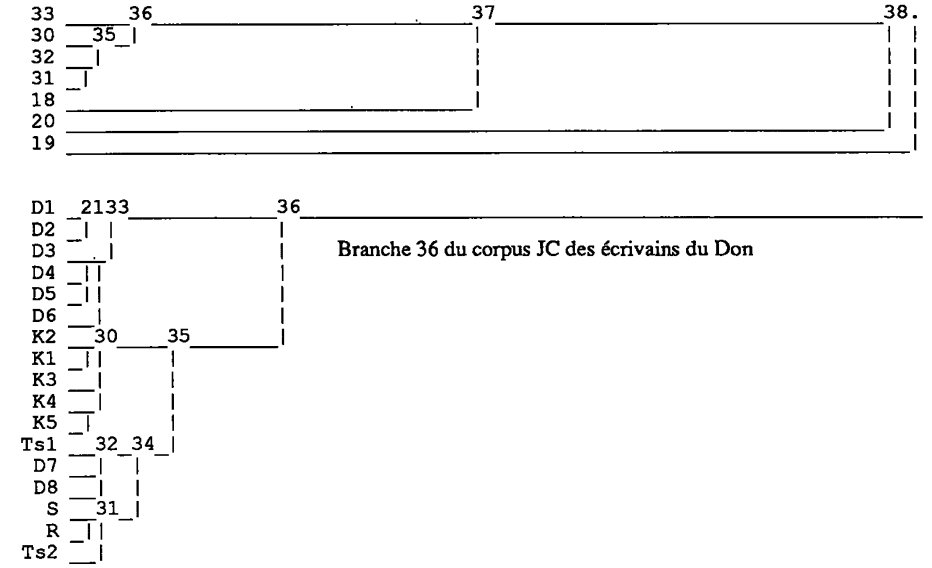
que la dispersion des fragments n'est pas prise en compte.

On voit ensuite, dans le plan (1, 3), que la bande du récit s'étale suivant l'axe 3, entre πusk, (F3>0); et JC, (F3<0). (C'est d'ailleurs principalement suivant l'axe 3 qu'au §4, JC se distingue des fragments de récits de JF).

N.B. Sur les plans, l'échelle de F2 et F3 est amplifiée relativement à F1.

```
en principal : IW x JC ∪ {Δerj, πusk, Léni} : 160 formes x 20 textes
trace  :   2.608e-1
rang   :      1     2     3     4     5     6     7     8     9    10
lambda :   1033   946   439    71    30    16    11     9     8     8 e-4
taux   :   3960  3626  1684   271   115    62    43    35    30    29 e-4
cumul  :   3960  7585  9269  9539  9654  9716  9760  9795  9825  9854 e-4
```

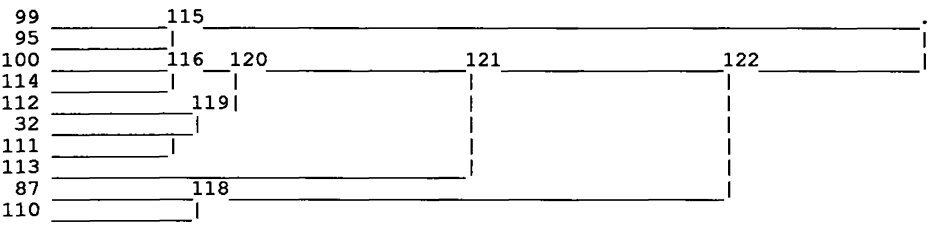
c	CAH(19) : Partition en 7 classes :	Sigles des textes de la classe c
33	D1 D2 D3 D4 D5 D6	le "Don Paisible", de M.A. Cholokhov
30	K2 K1 K3 K4 K5	les nouvelles de F.D. Kriukov
32	Ts1 D7 D8	œuvres diverses deM.A. Cholokhov
31	S R Ts2	
18	πusk	Pouchkine, représente le récit classique
20	Δerj	Derjavine, avec Pouchkine, représente la poésie lyrique
19	Léni	Lénine, représente l'exposé



Branche 36 du corpus JC des écrivains du Don

La CAH montre, pour l'ensemble des 20 colonnes principales, les trois genres: avec la structure bien connue intérieure à JC: KRIUKOV, à part: et les

c	Partition de JF en 10 classes : Sigles des fragments de la classe c	
99	ΔBog Δfa Δfb Lm2 Δiz2 πvs2 Δvd4 Δvd3	poésie de ΔERJAVINE ...
95	Δiz3 Δiz1 πvs1 Δvd1 Lm1	(manque Δvd2, *in* 112)
100	ß12a Kπa	
114	ß12c LMd Kπc KNy Kπâ KJa logb	prose de KARAMZINE ...
112	Δπb KJc Δπa KJb ßskf Δvd2 KJd Kπb †Re logc	
32	loga	
111	†Rb ß12b †Rd †Ra logß †Rf	†R : vie de ROUBLOV
113	LMc LMe LMb LGa LMa LGe LGc LGb LGd	LÉNINE
87	πB3 πB4 πB5	récit de POUCHKINE ...
110	πBb πBa πB1 πB7 πB9 πB2 KNz †Rc πB8 †RL πB6	



N.B. Dans la CAH ci-dessus les fragments de JF sont adjoints en supplément à l'analyse, objet du §5.

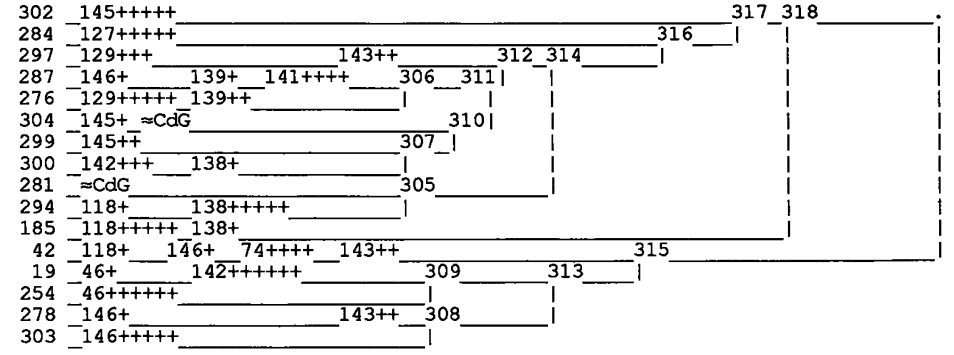
6 Appendice: Classification du lexique IW

Afin d'expliquer par quelles associations se constituent les classes de textes, on a fait une classification de l'ensemble IW des formes; et étiqueté, respectivement, chacune des deux partitions retenues pour IW et (JF ∪ JC) en fonction des classes de l'autre. À l'intention des lecteurs connaissant la langue russe ces résultats complémentaires sont ici publiés en appendice.

En haut du tableau de la CAH des formes, on a la classe i302, associée à la classe j145 des récits. Vient, ensuite, agrégée à i302 à un niveau très élevé, la classe i284, caractéristique de j127, qui comprend quatre fragments de LÉNINE (LM = *Matérialisme et Empiriocriticisme*); tandis qu'au bas du tableau, on trouve i303, associée à j146, où trois fragments du même LÉNINE (LMa et aussi LG = *l'État et la Révolution*) vont avec d'autres textes ayant rapport à la philosophie.

Dans la partition de IW retenue (définie par les 15 nœuds les plus hauts), certaines classes sont réduites à un ou deux mots. Ainsi, i185, там где, {là, où} va avec des fragments d'odes de DERJAVINE ou de LOMONOSOV, j118. Le mot вперед, {en avant}, est répété dans la fin du poéme de BLOK, *les douze*, ß12c; tandis que кругом та, sont dans le début de ce même poème. Enfin, или, {ou}, est dans un fragment, Δπb, où DERJAVINE explique, en prose, sa poétique. Avec un corpus plus étendu, de telles associations s'estomperaient.

c	Partition des 160 formes de IW en 16 classes : formes de la classe c
302	а уж тоже потом со куда у же так это вот ну опять почему
	хоть такая да свои были была
284	самой будто есть против себе
297	будут которой какие этим ни свое той часто иной точно бы
287	много несколько скоро таким почти этот своего
276	дома давно другой перед никого своему свою должно вместе своим
304	свое было уже один стал был ничего какая эта теперь своем всегда всего
	одна чтобы том тут все после чего об ко совсем не по как стало сам
	какой из к того от этом чем быть что через мало одной всех за чуть над с на
299	себя откуда кого эту весь до то сквозь одно вдруг чтоб
300	ли когда если лучше будут тогда нет
281	в во но и
294	кто тот под своих всем одну о вокруг свой лишь
185	там где
42	или
19	вперед
254	кругом та
278	даже которые очень всю этой такой при этого между тем
303	таких самого такое этих которую нельзя сейчас такие еще без только можно для



N.B. L'étiquetage de la partition du lexique IW des 160 formes, renvoie à la partition de l'ensemble (JF ∪ JC) des textes qui est publiée au §4.

7 Conclusion

À la différence de ce qui est pour le grec (cf. ce cahier, [DISCR. CAH LING.]) nous n'avons pas encore, en langue russe, saisi de corpus étendu. Mais il apparaît déjà qu'un lexique tel que IW, qui compte 160 formes de mots outil, peut servir à une taxinomie générale des textes en langue littéraire.

La diversité des genres est bien comprise; même si l'on s'interroge sur la longueur optima à donner aux fragments au sein desquels les formes sont dénombrées: s'ils sont courts, le profil en est imprécis; trop longs, ils absorbent les nuances.

Plus secrète est la signature de l'auteur.

Poète lyrique capable d'égaler DERJAVINE dans l'Ode; vif et prenant dans le récit plus qu'aucun prosateur; sentencieux, à ses heures, quand il critique ou entreprend de faire œuvre d'historien; POUCHKINE a mis le comble à la surprise des changements de décors et de costumes dans ce "*Boris Godounov*", où, parmi les échos de tous les genres, on ne cesse pourtant d'entendre son inimitable voix.

Mais l'axe sur lequel s'élève POUCHKINE est perpendiculaire à ceux que le statisticien a su extraire.

Références bibliographiques

[COMPAR. RUSSE]: G. et A. VOLOCHINE, "Étude comparée de textes russes: le Don Tranquille et d'autres œuvres de M.A. CHOLOKHOV; les nouvelles de F.D. KRIUKOV", in *CAD*, Vol. XX, n°1, pp. 7-26; 1995.

[DISCR. CAH LING.]: J.-P. & F. BENZÉCRI, "Analyse discriminante et classification ascendante hiérarchique dans l'adjonction d'individus à un échantillon de référence: application à des données linguistiques", in *CAD*, Vol. XX, n°1, pp. 45-66; 1995.

[TEXTES RUSSES]: J.-P. BENZÉCRI, "Sur l'étude des textes russes d'après les occurrences des formes de mots", in *CAD*, Vol. XIX, n°1, pp.7-34; 1994.

Remerciements: *Les auteurs remercient* Mr. et Mme G. et A. VOLOCHINE *dont les recherches de* [COMPAR. RUSSE], *sont à la base du présent travail.*