LES CAHIERS DE L'ANALYSE DES DONNÉES

G. VOLOCHINE
A. VOLOCHINE

Typologie de textes russes d'après remploi des parties du discours dans les trois mots initiaux de chaque phrase : applications à des œuvres relatives aux Cosaques du Don

Les cahiers de l'analyse des données, tome 22, n° 4 (1997), p. 421-442

<http://www.numdam.org/item?id=CAD_1997__22_4_421_0>

Les Cahiers de l'Analyse des Données
Vol. XXII - 1997/8 - n°4 - pp.421-442.

TYPOLOGIE DE TEXTES RUSSES D'APRÈS L'EMPLOI DES PARTIES DU DISCOURS DANS LES TROIS MOTS INITIAUX DE CHAQUE PHRASE: APPLICATIONS À DES ŒUVRES RELATIVES AUX COSAQUES DU DON

[TYP. INIT. PHRASE]

*G. & A. VOLOCHINE**

0 Objet de l'étude

Un précédent travail, (cf. [COMPAR. RUSSE] in *CAD*, Vol XX n°1, 1995), basé sur le dénombrement de mots-outil, a mis en lumière des différences notoires aussi bien au sein des œuvres mêmes de M. A. CHOLOKHOV (М. А. Шолохов): Premiers livres du Don Paisible et œuvres ultérieures; que dans un corpus où, à ces œuvres, étaient adjointes celles de F. D. KRIUKOV (Ф. Д. Крюков).

De façon précise, les œuvres analysées de M. A. CHOLOKHOV étaient: "Le Don Paisible" (Тихий Дон); "Ils ont combattu pour la patrie" (Они сражались за родину); "Les Terres Défrichées" (Поднятая целина) "Le Destin d'un Homme" (Судьба человека); et celles de F.D. KRIUKOV , cinq nouvelles: "La mère" (Мать); "Le journal d'un instituteur" (Дневник учителя); "Au pays" (В родном краю); "La cosaque" (Казачка); "La lame de fond" (Зыбь). Il s'agit de textes mettant en scène le milieu cosaque du DON et s'étendant de la Grande Guerre (1914-18) à la collectivisation, en passant par la guerre civile.

Nous avions été conduits à admettre que, d'une part, F.D. KRIUKOV n'avait pas écrit, même partiellement, le Don Paisible (contrairement à ce qu'affirmait un parti de polémistes tenaces); mais que, d'autre part, il y avait une forte probabilité pour que les six premières parties du Don Paisible n'aient pas été écrites (en totalité ou partiellement) par M. A. CHOLOKHOV. Au terme de notre travail, nous nous étions référés à des études (cf. Geir KJETSA et coll., bibliographie, *in fine*) sur l'identification d'auteurs par l'analyse de

(*) G. Volochine: Agrégée de Russe; A. Volochine: Ingénieur.

Les cahiers de l'analyse des données 0339-3097/97/04/© Elsevier, Paris

l'appartenance grammaticale des mots utilisés au début et à la fin de chaque phrase. L'occasion s'offrait ainsi de mettre nos conclusions à l'épreuve d'une nouvelle analyse, différente et complètement indépendante de la première. C'est en se basant sur ce principe que la présente étude reprend l'analyse des mêmes textes que dans [COMPAR. RUSSE], en ne considérant que les trois premières formes graphiques de chaque phrase. [Toutefois, au §5.2, sont pris en compte simultanément les données des deux types].

1 Élaboration de normes grammaticales

1.1 Principe de la caractérisation du style et normes pour le dénombrements des types de débuts de phrase

La présente étude est basée sur le fait que chaque auteur (comme chacun de nous d'ailleurs), a sa façon personnelle de commencer et de terminer une phrase, aussi bien en écrit qu'en parole. Cette "signature" dépend, certes, de plusieurs facteurs. Elle ne se manifeste pas obligatoirement à chaque phrase; mais elle est statistiquement présente dans un discours. C'est ce que nous avons voulu établir en analysant l'appartenance grammaticale des trois premiers mots graphiques de chaque phrase. L'étude des mots de fin de phrase se fera ultérieurement.

En bref, le tableau croise un ensemble de fragments avec un ensemble de triplets de catégories grammaticales. Le titre du présent article parle de "partie du discours", terme par lequel les linguistes désignent les grandes catégories de: {nom, verbe, adjectif, adverbe, préposition,...}. Ainsi on dira que k(tr, *f*) est le nombre des phrases du texte, ou fragment, *f*, commençant par le triplet tr=PrCdS2: {préposition, cardinal, nom à un cas autre que le nominatif}. On voit que pour caractériser le style, il a fallu distinguer des subdivisions au sein de chaque catégorie: et c'est de ce travail préliminaire qu'a dépendu le succès des analyses; cf. *infra*, §1.2.

La constitution d'une norme a présenté de nombreuses difficultés. Pour établir la liste des catégories grammaticales à prendre en considération, il a fallu chercher un compromis entre le désir d'être le plus précis possible et celui d'avoir des données statistiquement valables. Car la multiplication des catégories, en diminuant le nombre d'occurrences, peut noyer l'information recherchée dans un bruit inextricable.

Le nombre de catégories grammaticales une fois arrêté, il reste à assigner à chaque forme graphique une catégorie et une seule. Or, pour certaines formes, les diverses grammaires ne s'accordent pas entre elles. Ainsi en est-il, en particulier, dans le classement des particules et des adverbes; des adverbes utilisés comme prépositions; des conjonctions susceptibles d'être employées comme particules ou adverbes. Nous avons choisi le système qui a paru le plus logique, compte tenu de notre objectif. Et, de plus, lors de l'enregistrement, il a fallu prendre en considération le sens, certains mots pouvant se classer dans deux catégories différentes.

Mais d'abord, il faut définir ce qu'est une phrase. Nous avons adopté la définition suivante: une phrase est un ensemble de mots ou de signes graphiques, délimités par les points, les points-virgules, les points d'exclamation, les points d'interrogation, les points de suspension. Les tirets et les virgules, dont l'usage en russe est particulier, ont été considérés comme des formes graphiques et inclus dans le codage au même titre que les mots proprement dits. Dans les cas où une phrase contient moins de six mots graphiques la priorité a été donnée à la construction du début de phrase, la fin de phrase (dont il n'est d'ailleurs pas question dans les analyses rapportées ici) n'étant pas prise en compte.

En définitive, on a considéré 16 classes de formes graphiques, certaines étant subdivisées.

Nous devons mentionner que cette norme dérive de celle, plus complexe (en 23 classes) utilisée par Geir KETSAA et Bengt BECKMAN dans leur étude sur le "Don Paisible".

1.2 Les classes grammaticales adoptées et leurs symboles

Le classement décrit ci-après peut paraître hybride et arbitraire. En fait, il est basé sur des catégories généralement admises dans la plupart des ouvrages spécialisés. D'autres normes seraient possibles, à condition qu'elles ne soient pas en désaccord avec les classements admis; et fassent la preuve de leur intérêt stylistique. Pour la commodité du lecteur, l'inventaire ci-après est donné dans l'ordre alphabétique des symboles, de deux caractères, que nous avons adoptés. À ces symboles, en sont adjoints d'autres, d'un seul caractère, moins explicites, mais commodes pour les listages et graphiques.

Adjectifs: symbole: **Ad.** (@)

Adverbes: 3 catégories:

A1: (A) Adverbes de forme adjectivale: "медленно"; et les comparatifs: "медленне";

A2: (a) Autres adverbes et certains gérondifs considérés généralement comme adverbes: "молча"- "стоя" - "сидя";

Ap: (ã) Adverbes considérés comme issus de pronoms: "везде" - "всюду" - "всегда" - "здесь" - "иногда" - "тут" - "туда".

Conjonctions: 4 catégories:

C1: (C) Conjonctions simples (coordination, subordination): "но" - "кабы" - "если" - "чтобы";

C2: (Ç) Conjonctions ≈ pronoms:"каков" - "какой" - "кто" - "чей";

C3: (c) Conjonctions ≈ particules: "а" - "будто" - "ведь" ;

C4: (ç) Conjonctions ≈ adverbes: "даром" - "затем" - "пока".

Gérondifs: 2 catégories:

G1: (G) gérondifs présents: "читая";

G2: (g) gérondifs passés: "прочитав".

Cardinaux: Nb. (q)

Pronoms: 3 catégories:

P1: (i) Pronoms personnels, y compris les génitifs: "его" - "её" - "их", employés comme possessifs de la 3-ème personne, ainsi que le réfléchi: "себя";

P2: (j) Pronoms susceptibles d'être utilisés comme particules: "всё" - "всё-таки" - "себе" - "то-то" - "это";

P3: (k) Autres pronoms, y compris "тот" décliné (sauf "то"); et: "это".

Prépositions: 2 catégories:

Pr: (p) Toute préposition (,sauf les prépositions ≈ adverbes): "на" - "в" - "до" - "за";

Pp: (P) Prépositions ≈ adverbes: "возле" - "вокруг" - "вблизи" - "мимо" - "около".

Particules: 2 catégories:

Pa: (π) Particules généralement admises comme telles: "бы" - "ну" - "нет" (dans le sens de: "non");

Pd: (Π) Particules parfois classées comme adverbes:"ещё" - "почти" - "уже" - "уж" - "просто" - "прямо" - "некогда" - "никогда".

Participes: Pt. (é)

Négations: Pn: (N) "не" - "ни" - "нет" (dans le sens de: "il n'y a pas").

Substantifs: 2 catégories:

S1: (S) Substantifs au nominatif: sujet, attribut, interpellation et équivalent vocatif :"Господи";

S2: (s) Autres substantifs (sans distinction: animé-inanimé), l'analyse se faisant selon le cas grammatical et non selon la forme (qui peut être la même pour le nominatif et pour l'accusatif): "он положил книги (accusatif pl.; et non nominatif) на стол (acc.)".

Tirets: Tr. (-)

Verbes: 2 catégories:

V1: (V) Verbes imperfectifs: "читать" - "писать" - "входить";

V2: (v) Verbes perfectifs sous toutes leurs formes: "прочитать" - "написать" - "войти".

Virgules: **Vr**: (:) virgules, et autres signes de ponctuation {point; virgule; deux points} qui ne sont pas reconnus par nous comme étant des séparateurs de phrase (cf. *supra*, §1.1): {"," ";" ":"}.

Hybrides: Deux classes comprenant des conjonctions, éventuellement issues de pronoms et susceptibles d'être utilisées également comme particules, prépositions ou adverbes:

Xy: (y) {Conjonction + adverbe + particule}, {conjonction + adverbe + préposition}: "едва" - "либо" - "лишь" - "разве";

Xz: (z) {Conjonction + pronom + adverbe}, {conjonction + pronom + particule}: "то" - "что" - "где" - "когда".

2 Des données aux analyses

2.1 Codage d'un texte: attribution de sigles aux débuts de phrase

La norme étant choisie, tout texte peut être codé en attribuant à chaque phrase un sigle (de six caractères) formé des symboles attribués à ses trois premiers mots.

Voici quelques exemples de codage.

PrS2C3 (psc): pour: "из расположения ли ...": Préposition - Substantif (non nominatif) - Conjonction ≈ particule;

P1V1P3 (iVk): pour: "он обезоруживает моё ...": Pronom - Verbe imperfectif - adjectif Possessif;

PrS2V1 (psV): pour: "из коридора вторгается ": Préposition - Substantif (non nominatif) - Verbe imperfectif.

Ainsi, tout texte peut être traduit en une suite de sigles, qui en représentent les phrases.

2.2 Ensemble des 12 textes, ou fragments d'œuvres, ou recueils

Ainsi qu'on l'a annoncé dès le §0, le corpus des textes est essentiellement le même que dans [COMPAR. RUSSE]. Mais le choix des fragments est différent; car, en bref, on prend en compte, ici, non des mots, mais des phrases, dont le nombre est de beaucoup moindre. On a donc retenu un ensemble de 12 textes (ou fragments, ou recueils). Ces textes sont énumérés ci-après, par œuvres; avec, pour chacun, un symbole, suivi du nombre des chapitres (cap) et de celui des sigles (phrases) retenues.

{**D1**, ..., **D8**; **TS1**, **TS2**; **K**; **R**}.

"**Le Don paisible**": {**D1**, ..., **D8**}

Livre I (1928): **D1** (1-ère partie: 23 cap; 2520 sigles); **D2** (2-ème partie: 21 cap; 2817 sigles); **D3** (3-ème partie: 24 cap; 3877 sigles);

Livre II (1929): **D4** (4-ème partie: 21 cap; 3684 sigles); **D5** (5-ème partie: 31 cap; 4586 sigles);

Livre III (1933): **D6** (6-ème partie: 65 cap; 9783 sigles);

Livre IV (1940): **D7** (7-ème partie: 29 cap; 6924 sigles); **D8** (8-ème partie: 18 cap; 5118 sigles).

"**Les Terres Défrichées**": {**Ts1**, **Ts2**}:

Tome 1 (1932): **Ts1** (40 cap; 8731 sigles);

Tome 2 (1959): **Ts2** (29 cap; 8981 sigles).

"**Ils ont combattu pour la Patrie**" (1956-57): **R** (26 cap; 4751 sigles).

"**Nouvelles**", de F.D. KRIUKOV: **K** (cinq nouvelles, datant de 1896 à 1913, dont le recueil est considéré comme un texte unique; 5804 sigles).

Le texte: "Le Destin d'un Homme", de M. A. CHOLOKHOV, n'est pas pris en compte ici; car, (cf. [COMPAR. RUSSE], §1.4), ce récit à la première personne, devait manifestement se distinguer des autres textes du même auteur. quant à l'appartenance grammaticale des trois premiers mots de ses phrases.

2.3 L'ensemble des 92 sigles retenus: tableaux de correspondance

De même que, pour [COMPAR. RUSSE] (cf. §2.3), ne figurent dans le tableau analysé que les mots outil présents dans tous les textes; de même, ici, on a, d'abord, recherché tous les sigles attestés au moins deux fois dans le corpus; pour ne retenir, finalement, que ceux présents dans tous les textes. Il en est résulté un ensemble de 92 sigles ternaires; et un tableau (12×92), où (cf. *supra* §1.1) k(tr, *f*) est le nombre des phrases du texte ou fragment: *f*, commençant par le triplet: tr.

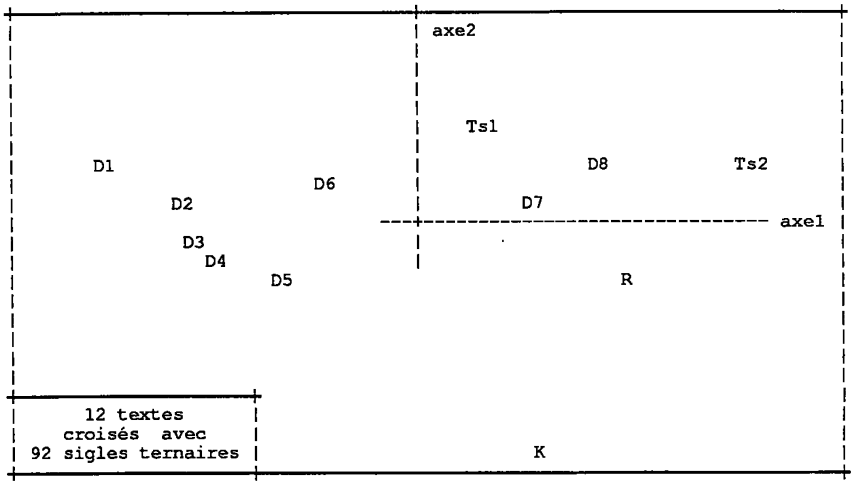
D'autre part, en vue d'apprécier dans quelle mesure la considération des deux premiers mots de chaque phrase pourrait suffire à caractériser le style, on a, semblablement, retenu un ensemble de 135 sigles binaires (i.e. formés chacun de deux des symboles grammaticaux expliqués au §1.2).

Aux §§3 et 5.1, on rend compte d'analyses factorielles et de classifications fondées, respectivement, sur les tableaux de correspondance: (92×12); et (135×12). Au §5.2 le dénombrement des mots outil est traité en même temps que celui des triplets.

3 Croisement de 12 textes avec 92 sigles ternaires

```
12 textes croisés avec 92 sigles ternaires
Trace :  1283 10-4
rang   :    1    2    3    4    5    6    7    8    9   10    11
Lambda:   468  173  124  104   84   75   64   59   53   42    36 10-4
Taux   :  3650 1350  970  810  660  582  500  464  412  327   282 10-4
Cumul  :  3650 5000 5970 6780 7440 8022 8522 8986 9398 9725 10000 10-4
```

On considérera successivement la représentation des ensembles des textes et sigles dans le plan (1, 2), issu de l'analyse factorielle; la CAH de ces ensembles; la correspondance entre classes de sigles et textes. Un commentaire stylistique fait l'objet du §4.

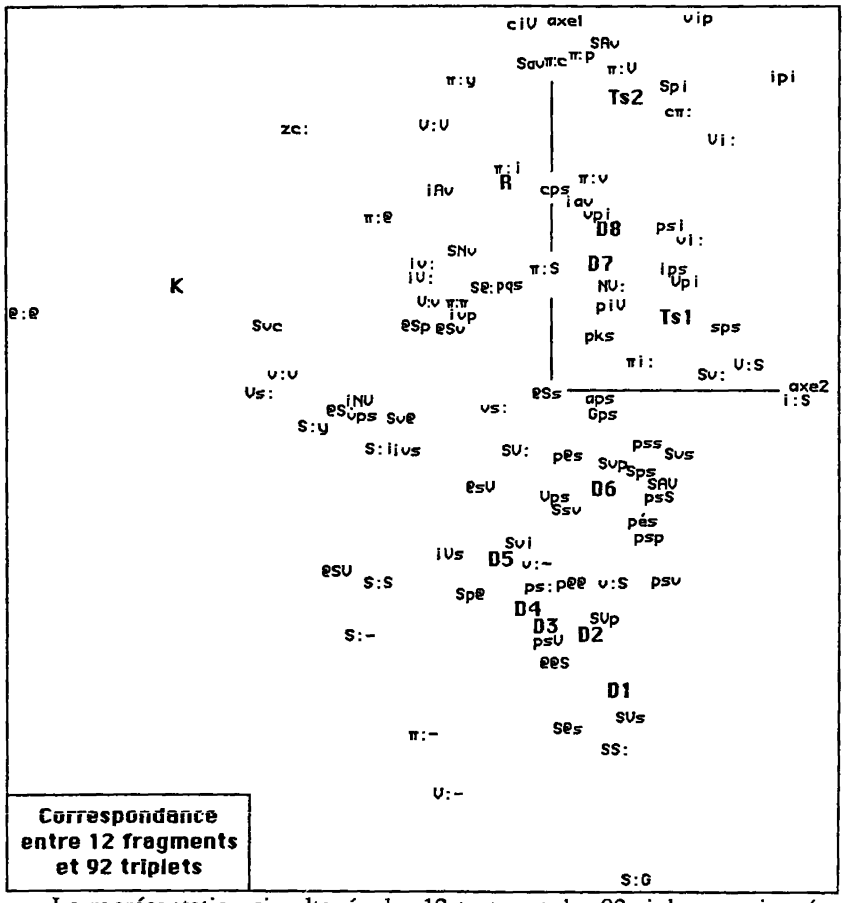


3.1 Analyse factorielle du tableau sigles × textes, (92 × 12)

On constate que les 3 premiers axes totalisent presque 60% de l'inertie, les deux premiers totalisant 50%. On se bornera à considérer le plan (1, 2); mais le rôle des facteurs F3 et F4 apparaîtra au §3.2, à propos de la CAH.

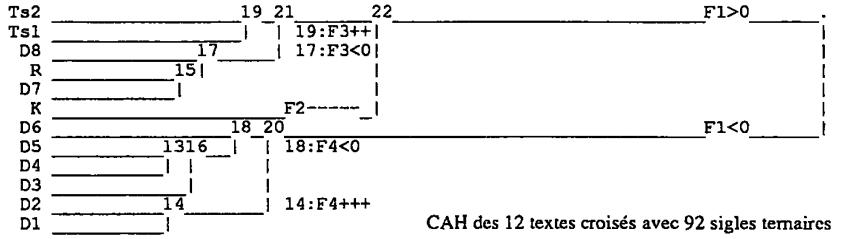
De même que dans [COMPAR. RUSSE], l'ensemble des textes, sauf ceux de F. D. KRIUKOV, se placent au voisinage du 1er axe. Les textes du Don Paisible s'étalent le long du 1er axe des valeurs négatives (D1 à D6) aux valeurs positives où D7 et D8 viennent se mêler au nuage des autres textes signé du nom de M. A. CHOLOKHOV (Ts1, Ts2, R). Les texte de F. D. KRIUKOV sont rejetés loin vers les valeurs négatives du 2ème axe.

Des images semblables de l'ensemble des textes ont déjà été vues au §3 de [COMPAR. RUSSE]. La seule différence étant que, dans la présente étude (cf. *supra*, §2.2), S ("Le destin d'un Homme") manque; et que l'on a dû cumuler dans K les cinq nouvelles {K1, ..., K5} de F. D. KRIUKOV.



La représentation simultanée des 12 textes et des 92 sigles ternaires (ou triplets) montre une image dense et qui serait inextricable si l'on n'avait réduit les sigles à 3 caractères.

On remarque toutefois que les sigles commençant par π (π=Pr=Particule) sont dans le demi-plan (F1>0), opposés aux premières parties du "Don"; fait seul exception π:– (':'=Vr=Virgule, ou autre signe; '–'=Tr=tiret). De plus, tous ceux de nos 92 sigles comprenant une particule (y compris π:π) commencent par π; la seule exception étant cπ: (exception relative, dans la mesure où c=C3=conjonction particule).



CAH des 12 textes croisés avec 92 sigles ternaires

3.2 Classification Ascendante Hiérarchique de l'ensemble des textes

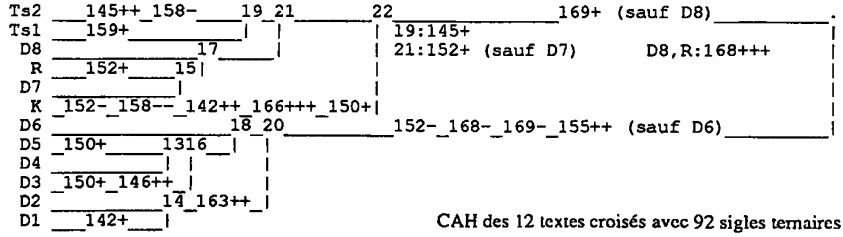
La CAH (effectuée dans l'espace des profils, rapporté aux 11 axes issus de l'analyse) partage l'ensemble des textes, suivant la direction de l'axe 1, en deux branches dont la première, '20', agrégée à un bas niveau, comprend les 6 premières parties du Don Paisible; le reste étant dans '22'. Au sein de '22', dans la direction (F2<0), les nouvelles de F. D. KRIUKOV, cumulées dans K, se séparent nettement de la branche '21' laquelle groupe les chapitres 7 et 8 du "Don Paisible" et tous les autres textes signés de M.A.CHOLOKHOV. Les deux tomes des "Terres Défrichées", {Ts1, Ts2}, très écartés vers (F3>0), constituent une subdivision, '19', de '21'.

À quelques variantes près, on retrouve, ici encore, la structure obtenue dans l'analyse par mots-outil (cf. [COMPAR. RUSSE], §4). Dans les deux CAH, {D1 ... D6} s'oppose au reste, au sein duquel K se détache nettement.

Mais, dans le partage de {D1 ... D6}, on voit {D1, D2} s'écarter nettement vers (F4>0) . Et, au sein du reste, une fois K mis à part, Ts1 rejoint Ts2 pour s'écarter vers (F3>0); {R, D7, D8} constituant une subdivision agrégée à un bas niveau.

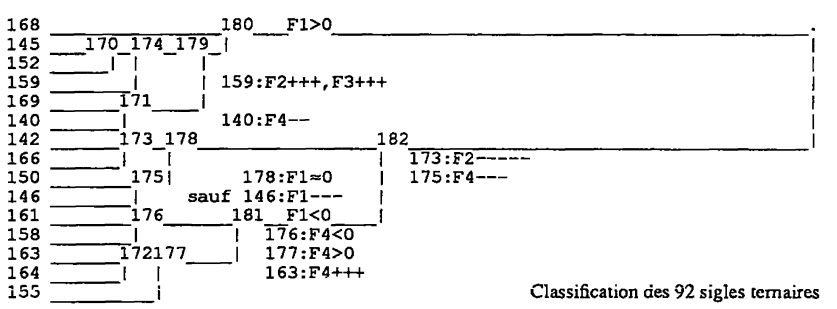
3.3 Classification Ascendante Hiérarchique de l'ensemble des sigles

On a retenu la partition en 15 classes définie par les 14 nœuds les plus hauts de la hiérarchie. C'est d'après cette partition qu'est étiquetée, ci-dessous, l'arbre de la classification des textes. Au §4, on considérera, de ce point de vue, les différences de style entre groupes de textes.



CAH des 12 textes croisés avec 92 sigles ternaires

c	Partition en 15 classes : Sigles des triplets de la classe c
168	SAv π:p Sav π:c zc: iAv iav
145	ciV vip cπ:
152	ipi π:V Spi Vi: vi: ips
159	sps πi: V:S i:S
169	π:y SNv π:v V:V cps π:S NV: pqs π:π @Sv π:i psı piV vpi S@: Vpi pks
140	iV: SV:
142	π:@ V:v @Sp vs: vps
166	iNV Vs: @S: Sv@ Svc v:v S:S @:@ S:y S:i S:-
150	ıv: ivp ıvs
146	@SV Ssv @@S iVs π:-
161	Sv: pss @Ss Sp@ p@s
158	aps Sps Svp Svs psS pés SAV psp
163	Gps Vps v:S SVp v:- ps:
164	@sV p@@ psv Svi SVs psV S@s
155	SS: V:- S:G



Classification des 92 sigles ternaires

Au sommet de la hiérarchie, l'ensemble des sigles se partage en deux branches, '180' et '182', dont les centres s'opposent suivant la direction du premier axe. Conformément à ce qu'on a vu dans le plan (1, 2), les sigles comportant une préposition, Pr=π, sont presque tous dans la branche (180: F1>0); avec deux exceptions: PrVrAd=π:@, qui est dans la subdivision (142: F1≈0); et PrVrTr=π:–, déjà vu, dans (146: F1<<0).

Il apparaît que les quatre sigles comportant un tiret (lequel est toujours au rang 3, et précédé d'une virgule ou autre ponctuation faible) sont dans la branche '182'; en revenant au plan (1, 2), on voit que ces sigles sont dans le quadrant (F1<0; F2<0).

classes de triplets x fragments (et cumuls de fragments)

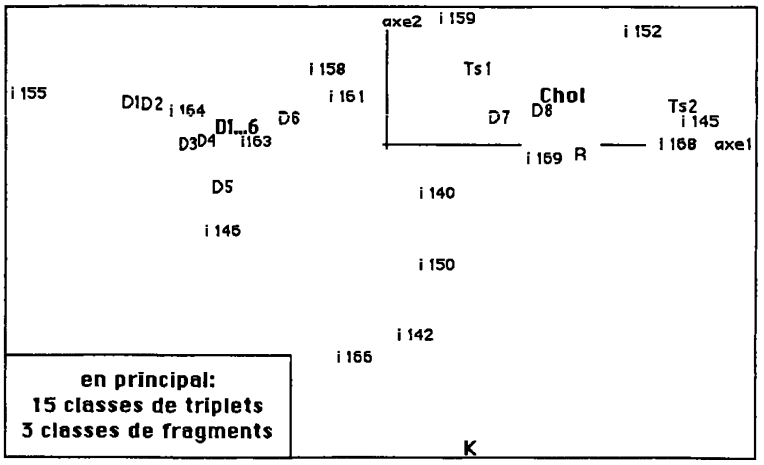
16014	D1	D2	D3	D4	D5	D6	D7	D8	Ts1	Ts2	R	K	D1..6	Chol
i168	39	44	45	52	75	163	162	190	145	210	141	122	418	848
i145	11	10	11	23	19	50	49	28	67	102	23	35	124	269
i152	31	46	70	42	53	133	115	104	196	186	106	58	375	707
i159	48	32	58	38	32	84	57	31	126	100	39	26	292	353
i169	132	154	180	162	223	468	462	265	471	456	243	305	1319	1897
i140	19	14	56	56	49	82	80	33	70	66	52	59	276	301
i142	60	43	43	38	38	109	81	50	67	88	39	104	331	325
i166	58	79	98	88	126	195	99	79	143	107	106	193	644	534
i150	16	26	79	34	80	110	61	68	106	100	32	93	345	367
i146	32	31	92	49	52	107	55	38	80	28	27	62	363	228
i161	61	124	130	152	166	352	191	124	277	197	107	117	985	896
i158	181	176	265	168	218	529	319	250	360	249	172	149	1537	1350
i163	121	150	129	98	121	227	140	98	143	137	84	101	846	602
i164	118	119	104	117	165	316	153	120	155	85	64	90	939	577
i155	62	64	81	73	78	98	38	20	74	35	22	29	456	189
Tot	989	1112	1441	1190	1495	3023	2062	1498	2480	2146	1257	1543	9250	9443

3.4 Correspondance entre classes de sigles et textes

Le tableau de correspondance entre fragments et classes de sigles permet d'apprécier l'intensité des contrastes signalé par l'analyse de données. Comme la CAH a confirmé le schéma ternaire du plan (1, 2), on a créé deux colonnes: D1...6, cumul des six premières parties du "Don"; et Chol, cumul du reste des textes signés de CHOLOKHOV.

En analysant, comme principal, le tableau à 15 lignes (classes de sigles) et 3 colonnes {K, D1...2, Chol}, avec en supplément les colonnes des fragments individuels, on n'obtient que deux facteurs; et l'on retrouve dans le plan (1, 2) la figure du §3.1.

De façon précise, sur l'ensemble des 12 fragments, la corrélation entre les facteurs 1 et 2 issus des deux analyses est respectivement: . 997; et .989



4 Commentaire stylistique

Le corpus est divisé en trois parties ou sous-nuages: {D1...6, Chol, K}, comptant repectivement {9250, 9443, 1543} occurrences des 92 triplets retenus pour l'analyse, objet du §3. Dans notre commentaire, on considère particulièrement des sigles, ayant une corrélation supérieure à 50% avec l'un ou l'autre des deux premiers axes; et qui selon leur place, caractérisent l'un ou l'autre des trois sous-nuages.

4.1 Nuage D1...6 des six premières parties du "Don"

Six sigles caractéristiques commencent par un substantif au nominatif (S1 = S = sujet, attribut ou interpellation), suivi, selon les cas, par:

un verbe imperfectif puis un substantif, non au nominatif (S1V1S2 = SVs); ou une préposition (S1V1Pr = SVp);

un autre substantif au nominatif et une virgule (S1S1Vr = SS:);

une virgule et un gérondif (S1VrG1 = S:G);

un adjectif et un substantif, non au nominatif (S1AdS2 = S@s);

une préposition et un adjectif (S1PrAd = Sp@).

Sur l'ensemble du corpus, le taux de ces six triplets S.. est (1256/20236); tandis que, pour D1...6, on a: (803/9250).

Le sigle PrS2V1 = psV = {préposition non adverbiale, substantif non au nominatif, verbe imperfectif} a pour taux repectifs: (278/9250) dans D1...6; et (476/20236) dans le corpus total.

À noter également les trois sigles {V1VrTr, V2VrTr, PaVrTr} = {V:–, v:–, π:–} dont la particularité est de se terminer par VrTr. Ensemble, ils ont pour taux: (220/9250) dans D1...6, et (367/20236) dans le corpus. En russe, le doublet VrTr est caratéristique de dialogues. Ainsi on a:

V1VrTr: "Продолжай, - (тебе говорю)"; "Жуёт, - (степенно ответил отец)";

V2VrTr: "Поняла, - (ответила она)"; "Поговорили, - (добавил он)";

PaVrTr: "А, - (отозвался голос)"; "Ну-ну, - (Степан присел на корточки)";

(dans ce dernier exemple, le premier mot est une interjection composée de deux syllabes; séparées par un tiret qu'on ne doit pas considérer comme une ponctuation). Dans certains cas on a une véritable conversation; dans d'autres, une simple amorce.

4.2 Nuage Chol = {D7-D8, Ts1/Ts2, R}

Nous notons deux groupes de sigles caractérisant cet ensemble de textes.

D'une part, 3 sigles en S1: {S1A1V2 = SAv, S1A2V2 = Sav, S1PrP1 = Spi}, constitués par un substantif au nominatif suivi soit par un adverbe et un verbe; soit par une préposition et un pronom. Avec pour taux: (532/9443) dans Chol; et (830/20236) dans le corpus total.

On notera la différence entre ceux des triplets commençant par S qui prédominent, respectivement, dans D1...6 et dans Chol. Dans Chol, si la phrase commence par un substantif au nominatif, elle se poursuit par un adverbe ou une préposition; et dans ce dernier cas on trouve en 3-ème position un pronom personnel (S1A1V2 = SAv, S1A2V2 = Sav, S1PrP1 = Spi). Alors que, dans D1...6, c'est plutôt un adjectif qui est choisi (S1PrAd = Sp@).

D'autre part, 4 sigles commençant par une particule suivie d'une virgule; le troisième terme pouvant être un verbe imperfectif, un pronom, une préposition ou une conjonction: {PaVrV1 = π:V, PaVrPr = π:p, PaVrP1= π:i, PaVrC3 = π:c}. Avec pour taux global: (507/9443) dans Chol; et (833/20236) dans le corpus. Cette structure est propre à des phrases de dialogue comme par exemple:

PaVrC3: "ну, а..."; "ох, да..."; "э, да..."; "может, и..."

PaVrV1: "ну, заводи..."; "небось, гляди..."; "ну, вот, думаю... "

PaVrPr: "может, через..."; "ну, с..."; "ей-Богу, без..."

Ces sigles attestent, dans Chol, une manière de rendre le dialogue, autre que celle déjà trouvée dans le groupe D1...6; l'une et l'autre manière étant rares dans K (nouvelles de F. D. KRIUKOV).

Et mention doit être faite du sigle PrS2P1= psi = {préposition, substantif non au nominatif, pronom}; lequel, quant au nombre des occurrences, est dans Chol, au 2ème rang après S1A1V2 = SAv. Les taux étant: pour psi (234/9443) dans Chol; et (409/20236) dans le corpus; et pour SAv (296/9443) dans Chol; et (456/20236) dans le corpus. Dans D1...6, si la phrase débute par une préposition et un substantif, le troisième mot est plutôt un verbe imperfectif (PrS2V1; au lieu de PrS2P1).

4.3 Le texte K, recueil des nouvelles de KRIUKOV

Une attention particulière doit être prêtée à deux sigles marquant une nette tendance à la description. Ce sont AdS1Vr = @S: (avec, en tête, un adjectif attaché à un substantif, généralement sujet de phrase); et AdVrAd = @:@ (deux adjectifs séparés par une virgule). Pour ce dernier, on citera: "красивый, белый (дом)", "сильная, бурая (лошадь)": deux adjectifs

habillent un même substantif. On trouve chez F. D. KRIUKOV jusqu'à trois adjectifs.

Ces séquences se rencontrent beaucoup plus rarement dans les deux autres groupes; D1...6 et Chol. Les taux sont: pour @S: (31/1543) dans K; et (216/20236) dans le corpus; et pour @:@ (24/1543) dans K; et (114/20236) dans le corpus.

On signalera encore: S1V2C3=Svc, taux (17/1543) dans K et (101/20236) dans le corpus. Ainsi que V1S2Vr=Vs:, taux (14/1543) dans K et (89/20236) dans le corpus.

5 Analyses complémentaires

Dans ce § on considérera brièvement, d'une part, au §5.1, un tableau analogue à celui objet du §3, mais prenant seulement en compte les deux mots initiaux de chaque phrase; d'autre part, au §5.2, un tableau dénombrant, à la fois, les occurrences de 92 triplets initiaux (cf. §3) et celles de 142 mots outil, pris en toute position (cf. [COMPAR. RUSSE]).

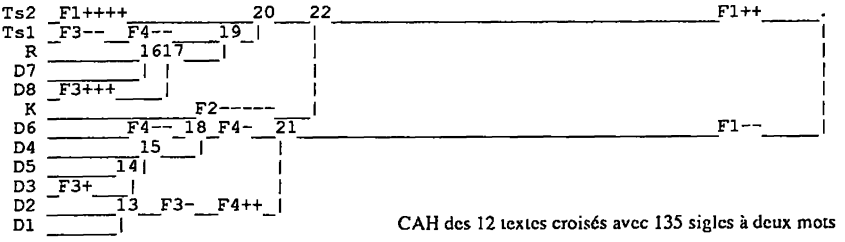
5.1 Croisement de 12 textes avec 135 sigles à deux mots

```
12 textes croisés avec 135 sigles à deux mots
trace  :   1.031e-1
rang   :    1    2    3    4    5    6    7    8    9   10   11
lambda :  442  125  102   87   74   43   38   35   33   29   23   e-4
taux   : 4282 1208  994  846  720  418  368  339  321  281  224   e-4
cumul  : 4282 5490 6484 7329 8049 8468 8835 9174 9495 9776 10000  e-4
```

Aux §§3.3, 3.4 et 4, on a souvent vu groupés des sigles ternaires commençant par les deux mêmes symboles. Il est légitime de chercher quelle information complémentaire apporte le troisième mot.

La norme et les symboles restant inchangés, on reprend ici l'analyse du §3, en ne retenant que les deux premiers mots de chaque phrase. De façon précise, on a 135 sigles à deux mots apparaissant au moins deux fois dans chacun des 12 textes analysés; d'où, par croisement, un tableau (135 × 12).

Le plan (1, 2) montre une disposition des textes très voisine de celle obtenue dans l'analyse avec les trois prèmiers mots. Et la CAH confirme le



CAH des 12 textes croisés avec 135 sigles à deux mots

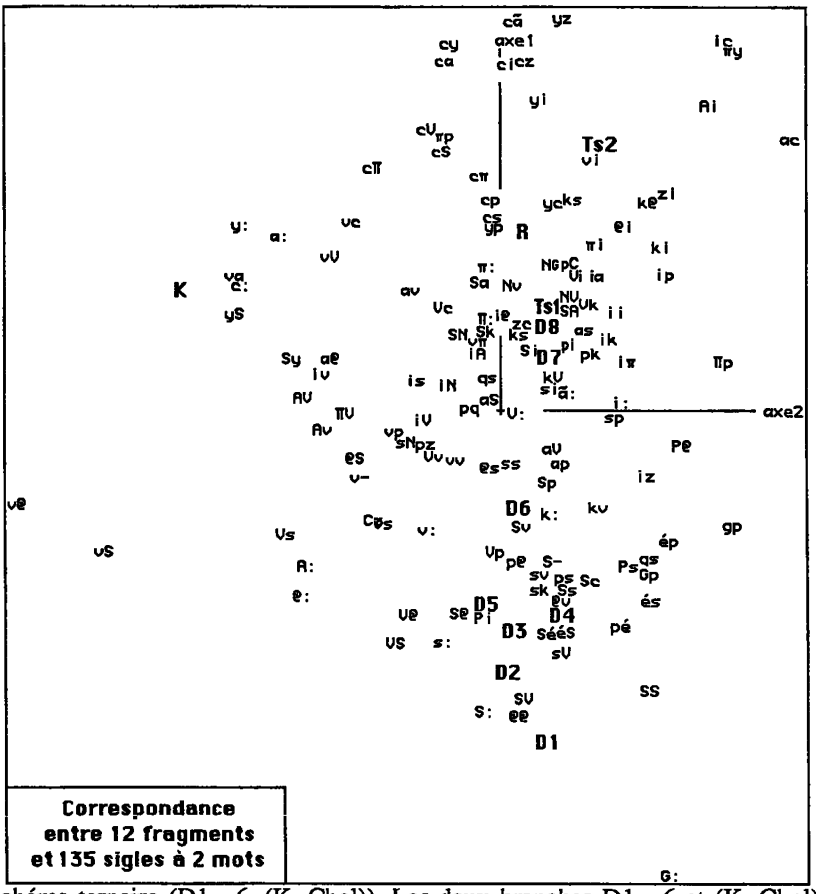
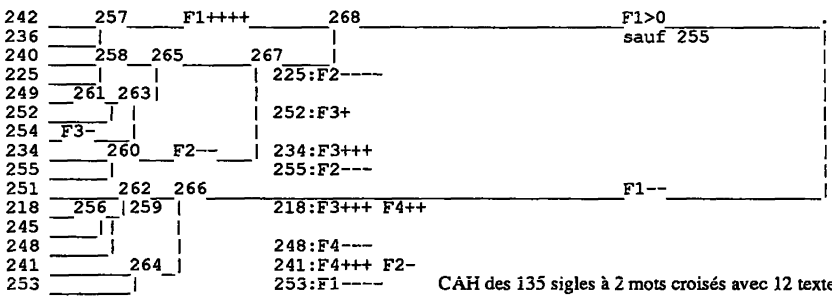


schéma ternaire (D1...6, (K, Chol)). Les deux branches D1...6 et (K, Chol) s'opposent suivant l'axe 1. La dénivellation relative entre les valeurs propres de rang 2 et 3 est moindre qu'au §3; mais l'axe 2 joue le même rôle, quasi exclusif, qui est de séparer K d'avec Chol; et les axes 3 et 4 n'interviennent que pour rendre compte des subdivisions au sein de Chol et de D1...6.

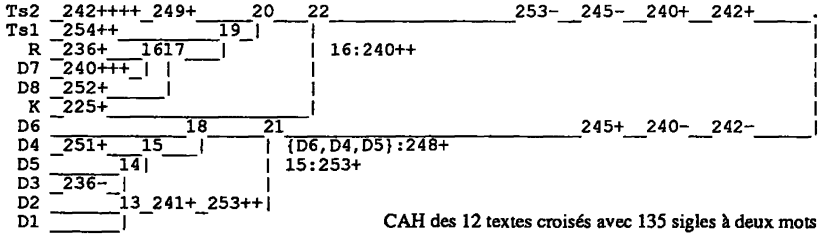
Comme au §3.2, dans D1...6, {D1, D2} s'agrègent à un bas niveau; et, au sein de D3...6, D6 s'oppose à D3...5. Dans Chol, {D7, D8, R} s'agrègent en une subdivision; mais, à la différence de ce qu'on a vu au §3.2, il n'y a pas de classe {Ts1, Ts2}; mais Ts2, isolé, s'oppose au reste de Chol.

c	Partition en 15 classes: Sigles binaires de la classe c
242	yz cã cz ci πp ic ac yi vi
236	cy ca cS c∏ cπ cp cs yp
240	cV π: Nv vc NG SN ki as NV
225	a: y: va vV c: yS AV
249	Ai πy @i zi ∏p P@ πi k@ Vk yc pÇ Vi ∏: Vc a@ kS
252	Sa SÃ i@ ia iA aS vv aV av vπ Si ik kV pk Sk iN pq pi ap
254	ks ip ii iπ ã: i: si is ∏V ss V: sp iz Sp
234	iv iV
255	Sy qs sN Çz Av @S v@ vS Vs V@ @:
251	zc Vv A: gp p@ @v
218	ép gs Gp Pi
245	Sv kv sv sk Ps ps pé és éS
248	@s k: S- Sc S@ s: sV Ss Sé @@
241	vp v- pz v: Vp vs VS
253	SV S: SS G:



CAH des 135 sigles à 2 mots croisés avec 12 textes

Comme au §3.3, on a fait une CAH des sigles; et étiqueté la classification des textes en terme de classes de sigles. Si l'on s'en tient aux nombres, il apparaît que les sigles à deux mots offrent une caractérisation des classes de



CAH des 12 textes croisés avec 135 sigles à deux mots

textes, et, notamment, de {D1...6, (K, Chol)}, équivalente à celle obtenue avec les sigles ternaires.

A première vue on pourrait donc se contenter de n'appliquer l'analyse qu'aux deux premiers mots.

Procédant comme au §4, nous avons considéré la distribution des sigles à deux mots ayant une corrélation supérieure à 50% avec l'un ou l'autre des deux premiers axes; et qui selon leur place, caractérisent l'un ou l'autre des trois sous-nuages. Mais si l'on tente de reprendre ici le commentaire linguistique, il s'en faut de beaucoup qu'on aboutisse à une synthèse aussi claire. La raison en est qu'un triplet initial évoque une phrase bien mieux que ne le fait une paire.

Au §4, on a vu que, dans Chol, si la phrase commence par un substantif au nominatif et une préposition, on trouve volontiers en 3-ème position un pronom personnel (S1PrP1). Alors que, dans D1...6, c'est plutôt un adjectif qui est choisi (S1PrAd). De semblables différences disparaissent ici.

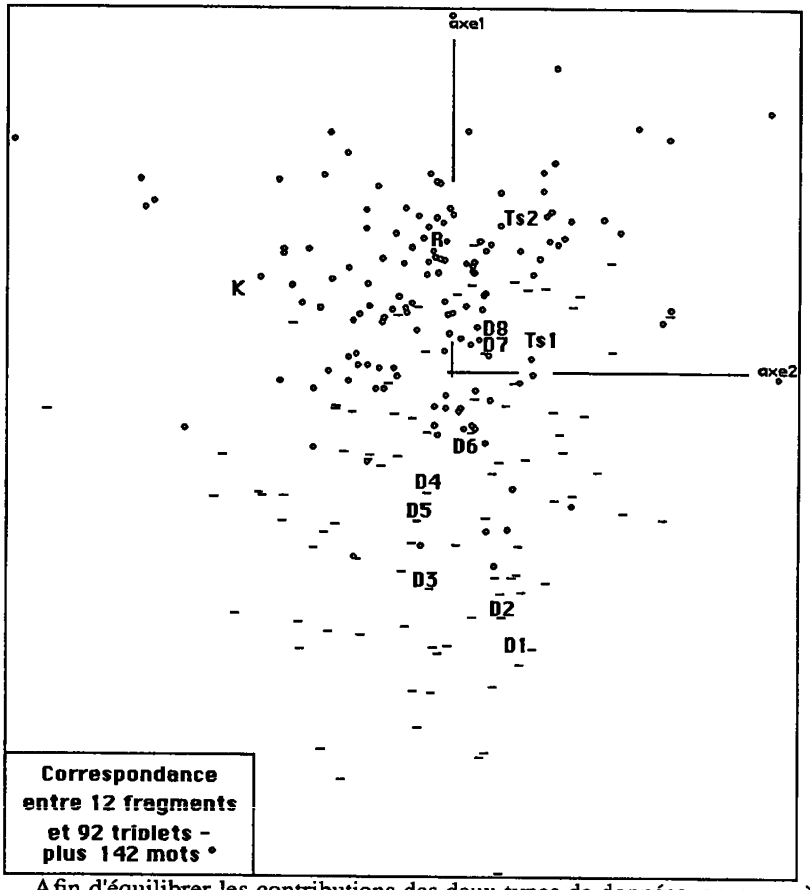
De plus, entre D1...6 et Chol, la forme du dialogue diffère; mais les triplets terminés par un tiret, caratéristiques de D1...6, ne laissent pas de trace si l'on ne considère que deux mots. Le tiret lui-même n'est alors compté que deux fois; dans (V2Tr = v-) et (S1Tr = S-).

Pourtant, répétons-le, les sigles à deux mots offrent une caractérisation numérique; passant par le cumul des fréquences suivant les classes de sigles. Mais, dans une fréquence de sigles à deux mots, joue, en quelque sorte, l'effet de multiples débuts ternaires qui pèsent inégalement selon le texte; et un cumul de fréquences de paires est encore moins évocateur; même s'il permet de décrypter la signature d'un auteur.

5.2 Croisement de 12 textes avec 92 triplets initiaux et 142 mots outil

```
12 textes croisés avec 92 sigles ternaires et 142 mots outil
trace : 8.392e-2
rang  :    1    2    3    4    5    6    7    8    9   10   11
lambda:  422   98   67   49   45   38   30   26   25   20   19 e-4
taux  : 5027 1166  804  579  540  451  353  313  294  241  232 e-4
cumul : 5027 6194 6997 7577 8117 8568 8921 9234 9527 9768 10000 e-4
```

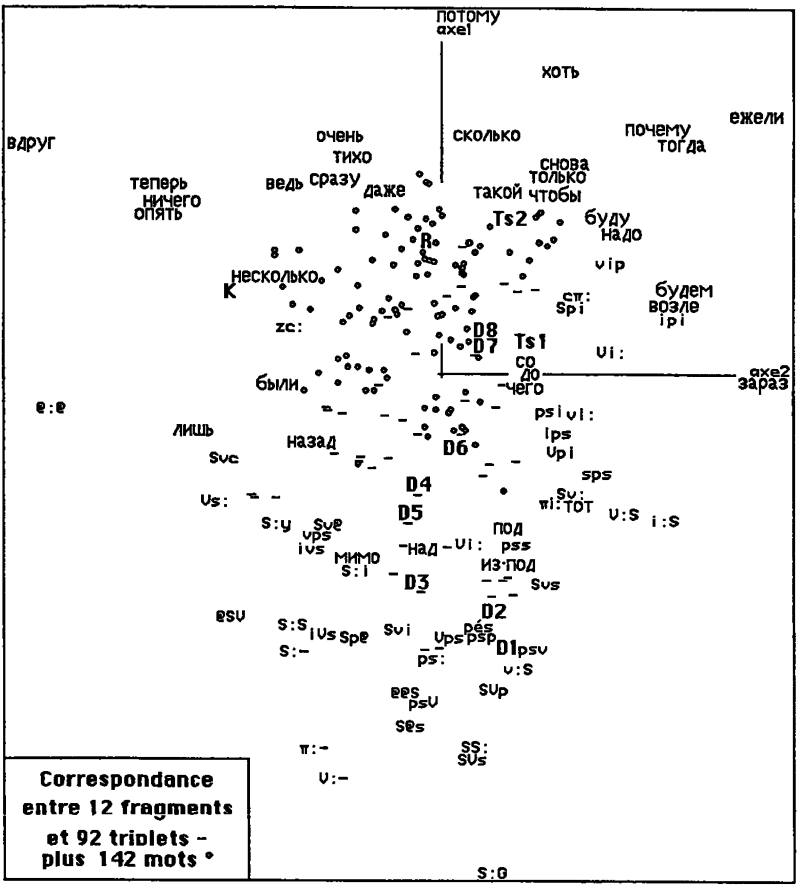
On analyse un tableau (12 × 234), dénombrant, à la fois, les occurrences de 92 triplets initiaux (cf. §3) et celles de 142 mots outil, pris en toute position (cf. [COMPAR. RUSSE]). Certes, les données ainsi assemblées ne sont pas parfaitement homogènes; mais toutes se réfèrent aux mots, pris tels quels et dénombrés en toute position; ou considérés quant à leur catégorie syntaxique et pris seulement en début de phrase. D'ailleurs en économie, en médecine..., on a traité avec fruit des tableaux encore moins homogènes que celui proposé ici.



Afin d'équilibrer les contributions des deux types de données, on pense à introduire des pondérations: car ainsi qu'on l'a dit au §2.2, un texte compte beaucoup moins de phrases que de mots. Dans l'analyse présentée ici, les nombres d'occurrences de mots sont gardés tels quels; et les nombres de sigles sont multipliés par 3.

Ce nombre pourrait être justifié parce qu'un triplet renferme des informations relatives à 3 mots distincts. Mais il a été, en fait, choisi par tâtonnement, en essayant les coefficients: 1, 3, 5, 10.

De façon précise, les tableaux ci-après donnent, pour 3 et 5 les taux



mots et triplets (*3)					
5	PDS	INR	CT1	CT2	CT3
Trp3	218	514	535	337	439
mots	781	488	456	659	554
Tot3	999	1002	991	996	993

mots et triplets (*5)					
5	PDS	INR	CT1	CT2	CT3
TRp5	312	594	586	464	558
mots	686	410	406	531	434
Tot5	998	1004	992	995	992

NB. dans ces tableaux, la ligne des totaux ne diffère de 1000 que par des erreurs d'arrondi

d'inertie cumulés dans l'espace des profils et sur chacun des axes 1 à 3.

La forme générale des résultats dépend peu du coefficient adopté. En bref, l'ensemble des textes se retrouve tel qu'on l'a toujours vu. Mais il est remarquable que les deux sous-nuages de variables, mots et triplets, sont

nettement décalés suivant l'axe 1; les triplets allant plutôt vers (F1<0), avec D1...6; et les mots vers vers (F1>0) avec Chol (et K).

Pour plus de clarté, on a donné du plan (1, 2) deux images. Dans l'une, les sigles et mots sont muets, ne figurant que par un caractère, ' - ' ou ' °'; ce qui montre le décalage des sous-nuages. Dans l'autre, on a, dans la mesure du possible, écrit en clair les sigles et les mots; ce qui permet notamment au lecteur d'apprécier dans quelle mesure, l'étalement des sigles suivant l'axe 1 reproduit, à un décalage global près, l'ordre vu au §3; et de même, pour les mots, dans [COMPAR. RUSSE].

6 Conclusions

Sur le choix du nombre de mots.

Peut-on se contenter d'une analyse portant sur les deux premiers mots; et non sur trois ?

Nous ne le pensons pas. Même si les graphiques des nuages et la CAH conduisent aux même conclusions dans les deux cas, il n'en est pas de même pour la caractérisation en termes linguistiques.

Sur les méthodes d'analyse

En se fondant sur la catégorie grammaticale des trois premiers mots de chaque phrase, on a retrouvé ici les mêmes résultats que dans [COMPAR. RUSSE], où l'on se fonde sur la fréquence d'utilisation des mots-outil dans le discours. Nous avons donc deux méthodes indépendantes qui se corroborent l'une l'autre.

Sur les résultats.

Nous pensons pouvoir affirmer, avec les précautions d'usage, que:

F. D. KRIUKOV n'a pas écrit le "Don Paisible", ni participé à l'écriture de cette œuvre.

M. A. CHOLOKHOV a certainement écrit "Судьба человека" (analysé dans [COMPAR. RUSSE] et laissé ici de côté). Il n'a probablement pas écrit les premiers livres du Don Paisible (D1, D2, D3).

Pour le reste des textes, il est raisonnable de penser qu'ils ont été écrits par M.A.CHOLOKHOV avec des inclusions plus ou moins importantes provenant d'un auteur inconnu contemporain de ce milieu cosaque, qui est si bien décrit au début du "Don Paisible".

Il reste à analyser les fins de phrases; et, si celles-ci sont porteuses d'information, faire une synthèse des débuts et des fins. Ce sera la prochaine étape.

Remerciements: Les auteurs remercient le Pr. J.-P. Benzécri qui a bien voulu s'intéresser au présent travail et en parachever la publication. Après mûre réflexion, celui-ci nous propose l'épilogue suivant.

«Le problème de la multiplicité des auteurs des textes d'un corpus renvoie à celui de la diversité de langue et de style que l'on peut trouver au sein d'un ensemble d'œuvres dont l'attribution à un auteur unique est incontestée.

«Or des auteurs qui, comme VOLTAIRE, en Français, ou POUCHKINE en Russe, ont brillé dans tous les genres, semblent couvrir sous leur manteau tout ce que leurs contemporains ont pu écrire. N'y a-t-il pas, à première vue, plus de similitude entre deux tragédies d'auteurs différents qu'entre un conte et une tragédie d'un même auteur ? Nous ne savons, encore, à ce niveau de généralité, en quoi consiste la *signature*.

«En restant dans le genre du récit, le problème se restreint; et un projet de recherche plus abordable se présente. Prendre, chez un auteur, une suite de volumes parus sur une période de dix ou vingt ans et composant un cycle unique. On pense aux "*Hommes de bonne Volonté*", de Jules ROMAIN.

«Ainsi, on revient au "Don Tranquille". Quant à cette œuvre, il est frappant que les huit parties s'ordonnent quasi rigoureusement, de D1 à D8, suivant l'axe 1 de toutes les analyses qu'on a pu faire du corpus; avec toutefois, de D6 à D7, une distance plus grande, qu'entre les autres couples de parties consécutives. Une telle continuité s'explique-t-elle par une mutiplicité d'auteurs qui se succèdent; ou plutôt collaborent suivant un mode progressif. Ou suffit-il d'évoquer un seul auteur qui mûrit, avec le renouvellement de ses personnages au sein d'une société qui bouleverse tous les rôles ?

«L'analyse du §5.2 a montré, sur l'axe 1, un décalage entre triplets, associés aux premiers livres du "Don", et mots, associés à des textes publiés plus tard. De ce décalage, qui nous a surpris, l'interprétation serait la suivante. Un texte compte, relativement, d'autant plus de mots et d'autant moins de triplets que ses phrases sont, en moyenne, plus longues. Quels que soient précisément ces mots et ces triplets, un gradient de longueur de phrases se manifeste par une opposition entre sigles et mots. Même si la longueur des phrases (notion, d'ailleurs incertaine, comme tout ce qui dépend de la ponctuation; laquelle ne s'impose pas toujours) ne suffit pas à ordonner validement les textes du corpus étudié, elle peut y contribuer.

«Tout en appréciant grandement ce que le sens linguistique des auteurs a permis de verser dans l'alambic de l'analyse des données; et les traits de structure incontestables qui se sont condensés sur les graphiques et les tableaux qui illustrent le présent mémoire; nous attendons la suite des travaux de G. & A. VOLOCHINE, pour arbitrer entre CHOLOKHOV et ses critiques. »

Références bibliographiques

Sur la langue russe:

Grammaire de l'Académie des Sciences de l'URSS; (1963);

Grammaire de la langue russe littéraire contemporaine; Académie des Sciences de l'URSS; (1970);

Dictionnaire de la langue russe; sous la direction de Д. Н. Ушаков; (1947);

Sur l'analyse des textes:

Geir KJETSAA, Sven GUSTAVSSON, Bengt BECKMAN, Steinar GIL: *The authorship of the Quiet Don*; Solum Vorlag (Oslo) & Humanities Press (N.J.); (1984);

Geir KJETSAA : "Storms on the Quiet Don"; *Pilot Study Scando-Slavica*; Vol XXII; (1976);

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"; [COMPAR. RUSSE]; in *CAD*; Vol. XX, n°1, pp. 7- 26; (1995);