L'étude des longueurs de phrases
Un réexamen des méthodes

Étienne ÉVRARD

0.0. Les deux premières publications du L.A.S.L.A. contiennent en appendice des remarques (à la rédaction desquelles j'ai largement participé) sur les longueurs de phrases, respectivement dans les *Consolations à Polybe* et *à Helvia* de Sénèque[1]. Il nous semblait, comme à de nombreux chercheurs, que la longueur des phrases était un caractère typique des textes littéraires et que l'étude devait en être fructueuse. Elle pouvait, à notre avis, se faire en deux directions diverses et, en un certain sens, complémentaires : tout d'abord, il est possible de déterminer la distribution globale des longueurs de phrases dans un texte et d'en tirer des indices tels que la longueur moyenne et la dispersion; par ailleurs, on peut aussi s'intéresser à la manière dont les longueurs se succèdent effectivement dans le texte; on a alors une étude pour ainsi dire chronologique susceptible d'être révélatrice à différents points de vue : la diversité éventuelle des parties d'une œuvre; l'emploi de mini-patterns dans de petits groupes de phrases;...

0.1. Ces appendices ont fait l'objet d'un article critique de M. Tore Janson, qui, tout en reconnaissant l'intérêt de l'entreprise du L.A.S.L.A., conteste le bien-fondé des méthodes utilisées[2]. Dès 1966, M. L. Delatte avait répondu à M. T. Janson dans

---

[1] L. DELATTE et Ét. ÉVRARD, *Sénèque - Consolation à Polybe : Index verborum : Relevés statistiques*, Liège, 1962, p. 195–202 et *Sénèque : Consolation à Helvia : Index verborum : Relevés statistiques*, La Haye, 1963, p. 257–265.

[2] Tore JANSON, "The problems of measuring sentence-length in classical texts", dans *Studia linguistica*, XVIII,1, 1964, p. 26–36.

un article de *Revue*[3]. Il m'a paru qu'après 25 ans, il ne serait pas inutile de revenir sur la question : des techniques nouvellement mises au point pourraient résoudre certaines des difficultés soulevées; par ailleurs, l'examen des tests statistiques et de leur validité mériterait d'être repris et approfondi.

*
* *

1. Dans l'évaluation des longueurs de phrases, nous avions pris le mot comme unité, malgré la difficulté de le définir de manière pleinement satisfaisante (difficulté dont il ne faut d'ailleurs pas surestimer les inconvénients pratiques). C'est l'unité qu'à ma connaissance adoptent tous les chercheurs qui se sont intéressés à la question. M. Tore Janson, dans une note, suggère que la syllabe serait peut-être préférable parce qu'on la définit plus aisément et que sa longueur varie moins[4], mais il ne met pas en pratique cette suggestion. Je ne crois dès lors pas utile de m'y attarder[5].

*
* *

2.0. Une autre critique, plus importante, porte sur la définition de la phrase. Nous avions décidé tout simplement de traiter comme limites de phrases les points, les points d'interrogation et les points d'exclamation[6], et de suivre fidèlement la ponctuation de l'édition Hermes, qui avait été choisie pour la mise en mémoire du texte.

2.1. M. Tore Janson s'interroge d'abord sur la valeur de la ponctuation que l'on trouve dans les éditions. De fait, il est notoire que la ponctuation des auteurs anciens est totalement due aux éditeurs modernes et que, pour un même texte, elle

---

[3] L. DELATTE, "A propos de la mesure des longueurs de phrases dans les textes classiques", dans *Revue - Organisation internationale pour l'étude des langues anciennes par ordinateur*, 1966, n° 1, p. 51–65.

[4] T. JANSON, *l.l*, p. 27, n. 3.

[5] On notera toutefois que F. RUSSO, "Utilisation de la mesure des longueurs de phrases pour la détermination de la structure de la prose de Fronton dans le *De Eloquentia* et le *De Orationibus*", dans *Revue*, 1975, n° 2, p. 31–48, utilise ce mode d'évaluation (p. 36), en le rapportant à J. Marouzeau, sans citer, sur ce point, T. Janson, que pourtant elle connaît.

[6] T. JANSON, *l.l.*, p. 28, n. 1, relève deux cas où cette règle n'a pas été appliquée : il s'agit de *Pol.*, 15, 1 et de *Helv.*, 12, 3. Dans ces deux passages, le point d'interrogation ou d'exclamation se trouve de toute évidence à l'intérieur d'une séquence syntaxique qu'il n'est pas possible d'interrompre; pour le français, ce genre de cas est indiqué dans GREVISSE, *Le bon usage*, 12e éd., par A. Goosse, 1988, §119, R. 2 et 122, R. 2.

varie d'une édition à l'autre. Il est cependant clair qu'elle reflète la perception que l'éditeur a de la syntaxe du texte et que cette perception, sauf cas extrêmes, a des chances d'être à peu près la même pour tous les éditeurs (et lecteurs) d'un même texte. On peut donc s'attendre à ce que les différences de ponctuation entre plusieurs éditions soient minimes et non significatives au sens statistique. C'est effectivement à ce résultat qu'est arrivé W. C. Wake, dans un article que nous citions et dans lequel sont comparées deux éditions des *Catégories* d'Aristote[7]. Depuis lors, un étudiant en philologie classique de l'Université de Liège, qui, par la suite, s'est tourné vers l'histoire des sciences et s'y est taillé une réputation enviable, M. Robert Halleux, a consacré son travail de fin d'études à une étude des longueurs de phrases dans l'*Octavie*, comparée de ce point de vue à deux tragédies de Sénèque, *Oedipe* et *Agamemnon*. Auparavant, il a pris la précaution de comparer, pour chacun des trois textes, la ponctuation de trois éditions : comme Wake, il a constaté que les différences n'étaient pas significatives; il note toutefois à juste titre que le caractère dialogué du texte réduit les possibilités de variations dans la ponctuation[8].

2.2. M. Janson considère que la recherche de Wake est de nature à clarifier le problème. Bien plus, il y relève une restriction qui lui paraît importante : d'après Wake, les divergences d'édition à édition se réduisent particulièrement si l'on traite sur le même pied le point et le point en haut[9]. C'est, pour Janson, l'occasion de discuter la décision que nous avions prise de ne pas tenir compte des points-virgules. Je crois qu'il y a là l'effet d'une confusion : il est vrai que le point-virgule, dans l'un de ses emplois, répond au point en haut, mais il a un autre emploi qui n'a pas de correspondant en grec, c'est celui qui isole un membre de phrase assez important mais ne constituant toutefois pas une séquence syntaxiquement indépendante. Cet usage est bien connu en français[10] comme en allemand[11]. On comprend sans peine que les éditeurs l'aient introduit dans les textes latins. Un exemple convaincant en est fourni par Suétone, *Domitien*, X, que je cite d'après l'édition de Roth, Teubner, 1904, en respectant sa ponctuation. Après le membre de phrase *discipulum... occidit* vient un point-virgule suivi de *item* et d'un accusatif avec ses déterminations, le tout dépendant toujours d'*occidit* : la portion de texte qui suit le point-virgule ne possède pas d'indépendance syntaxique; ce serait une erreur de la traiter comme une phrase. Un peu plus loin, on lit *Complures senatores... interemit; ex quibus Civicam*

---

[7] W. C. WAKE, "Sentence-length distribution of Greek authors", dans *Journal of the Royal Statistical Society*, series A, 120, 1957, p. 331–346.

[8] R. HALLEUX, *Octavie : Index verborum - Relevés grammaticaux - Etude sur les longueurs de phrases*, Liège, 1968, p. 299–311 (inédit).

[9] Cfr WAKE, *l.l.*, p. 334 et T. JANSON, *l.l.*, p. 28.

[10] Cfr GREVISSE, *Le bon usage*, §128 et 210.

[11] Cfr DUDEN, *Die Zweifelsfälle der deutschen Sprache*, éd. 1972, p. 593 s.

Cerealem ..., quasi molitores rerum nouarum. A moins de considérer *quibus* comme un "relatif de liaison", on peut difficilement voir dans le point-virgule une ponctuation forte équivalant à un point. Après *nouarum* vient un point-virgule suivi de *ceteros leuissima quemque de causa : Aelium Lamiam ob suspiciosos... iocos, quod... dixerat, quodque... responderat...; Saluium Cocceianum, quod Othonis... diem natalem celebrauerat; Mettium Pompusianum, quod... ferebatur et quod ...circumferret, quodque... indidisset; Sallustium Lucullum..., quod... passus esset; Iunium Rusticum, quod... edidisset appellassetque; cuius criminis occasione philosophos omnis urbe Italiaque summouit*[12]. Le mot *ceteros* et la série d'anthroponymes à l'accusatif sont régis par le verbe *interemit*; il en résulte que les expressions séparées une fois par un double point et les autres par des points-virgules ne constituent pas des entités syntaxiques indépendantes; on se tromperait grossièrement si on les considérait comme des phrases. En fait, il s'agit de membres de phrases que pourraient séparer des virgules, n'était le fait qu'ils sont assez longs et comportent déjà des divisions marquées par ce signe.

L'édition Hermes de Sénèque nous fournit des exemples analogues; ainsi, dans le *De Prouidentia*, I, 2, on lit : *superuacuum est in praesentia ostendere non sine aliquo custode tantum opus stare nec hunc siderum coetum discursumque fortuiti impetus esse...; non esse materiae errantis hunc ordinem...* Il est clair que ce qui suit le point-virgule est une infinitive régie par *superuacuum est ostendere*, au même titre que *stare* et que *hunc... coetum...esse*. On ne pourrait en faire une phrase distincte.

Il n'est cependant pas rare que le point-virgule sépare deux groupes syntaxiquement indépendants, qu'il y aurait lieu de traiter comme des phrases. En voici, toujours tiré de l'édition Hermes, un exemple de Sénèque, *Cons. ad Heluiam*, VII, 2 : *Atheniensis in Asia turba est; Miletus quinque et septuaginta urbium populum in diversa effudit; totum Italiae latus, quod infero mari adluitur, maior Graecia fuit.*

Dès lors, on voit bien la difficulté. La double valeur du point-virgule provoque nécessairement l'incertitude : ou bien, en suivant notre règle, on néglige un nombre indéterminé de phrases, ou bien, en considérant le point-virgule comme une limite de phrase, on prend en compte des groupes de mots qui n'ont aucune indépendance syntaxique.

On pourrait faire, pour le double point, des observations analogues[13]; on en a d'ailleurs vu un exemple dans le texte de Suétone qui vient d'être cité; il n'est sans doute pas nécessaire de s'y étendre plus longuement.

---

[12] Dans l'édition Ailloud, CUF, 1957, la ponctuation est la même sinon qu'après *nouarum*, il y a une simple virgule, mais qu'après *causa*, il y a un point, lequel conduirait à distinguer indûment une phrase supplémentaire.

[13] GREVISSE, §129 et 210.

Plus grave encore, il arrive que de simples virgules séparent des séquences qui, du point de vue de la syntaxe, sont tout à fait indépendantes[14]. Ainsi, dans Sénèque, *Cons. ad Polybium*, I (20), 1 : *aliud alio modo fragile est, rerum exitus uariantur, ceterum quicquid coepit et desinit.* Comme on le voit dans cet exemple, la virgule n'empêche pas l'indépendance syntaxique; elle note simplement l'existence d'une étroite connexion sémantique et logique.

2.3. Avant de chercher un remède à la situation qui vient d'être décrite, disons encore, pour répondre à une critique de T. Janson[15], que si, pour notre étude des longueurs de phrases, nous avions décidé de nous en tenir à l'édition de Hermes, c'est que nous pouvions ainsi traiter tout l'œuvre en prose de Sénèque dans un texte dû à un seul et même éditeur. Sans doute avons-nous conscience que la ponctuation de Hermes n'est qu'une approximation de la ponctuation idéale. Mais l'unité d'éditeur permet d'espérer une diversité minimale dans le mode de ponctuation.

2.4.0. Revenons à notre problème. La première tâche est de donner de la phrase une définition qui se fonde sur un ou des critères objectifs et ne laisse place à aucune indécision[16]. En suivant les principes de la syntaxe dépendancielle, tels qu'ils ont été énoncés par L. Tesnière[17], nous dirons qu'une phrase se compose d'un terme indépendant, qui en constitue le nœud principal, et de tous les termes qui, directement ou indirectement, sont régis par ce terme central. Il faudrait donc, sans trop se préoccuper de la ponctuation introduite dans les éditions de textes, décrire les chaînes de rection rattachées à un même nœud principal, déterminer les limites du bloc ainsi défini et les traiter comme limites de phrases.

2.4.1. Précisément, dans la communication que j'ai faite au 5e Congrès international de linguistique latine, j'ai proposé une méthode propre à traiter avec l'aide de l'informatique la relation de dépendance syntaxique[18]. Soit un texte saisi en mémoire d'ordinateur; chaque mot de ce texte peut être pourvu d'un certain nombre de descripteurs : parmi ceux que nous utilisons au L.A.S.L.A. se trouve le numéro d'ordre du mot dans le texte (ce numéro est calculé automatiquement); la méthode que j'ai proposée consiste à doter chaque mot d'un descripteur supplémentaire qui

---

[14] Cfr GREVISSE, §210, qui mentionne l'existence de cet usage de la virgule en français.

[15] T. JANSON, *l.l.*, p. 28.

[16] Je laisserai de côté la suggestion de L. DELATTE, *l.l.*, p. 62 s., reprise par F. RUSSO, *l.l.*, p. 32 s., de considérer comme une phrase ce qui est limité par une clausule : la théorie des clausules est chose assez incertaine; par ailleurs, les clausules peuvent se trouver ailleurs qu'en fin de phrase; enfin, l'usage des clausules n'est pas régulier chez tous les auteurs.

[17] L. TESNIÈRE, *Eléments de syntaxe structurale*, Paris, 1959.

[18] Ét. ÉVRARD, "Essai d'informatisation d'une syntaxe dépendancielle en latin", dans M. Lavency et D. Longrée éd., *Actes du 5e Congrès international de linguistique latine* (Louvain-la-Neuve, mars 1989) = *CIIL*, 15, 1989, p. 115-126.

n'est autre que le numéro d'ordre dans le texte du mot qui en est le régissant; le nœud principal ne peut recevoir de descripteur de ce type, et c'est ce qui permet de le repérer automatiquement. Ce n'est pas le lieu de revenir sur tout ce que l'on peut tirer d'un fichier enrichi de cette manière. Je rappellerai seulement qu'un programme d'ordinateur que j'avais présenté au Congrès permet de constituer automatiquement l'arborescence liée à un nœud donné, de calculer le nombre des niveaux de dépendance qu'il régit et de compter les éléments (ou mots) qui le constituent.

Dans la mesure où les numéros de régissants ont été correctement choisis et enregistrés, on a là un moyen de déterminer automatiquement, sur un critère rigoureux, les limites et la longueur des phrases. Il faut dire que, dans la préparation des fichiers, un certain nombre de points doivent au préalable être précisés; il en est ainsi, par exemple, du relatif de liaison, dont la réalité paraît pleinement assurée à la plupart des grammairiens mais dont certains linguistes contestent complètement l'existence tandis que d'autres pensent qu'il ne fonctionne que dans certaines conditions limitées[19]. Il n'en reste pas moins qu'une fois connues les règles que l'on s'est assignées, on a l'assurance d'une découpe en phrases parfaitement cohérente. Il semble donc qu'au fur et à mesure que des fichiers auront été traités en vue de l'étude de la syntaxe dépendancielle, le problème des longueurs de phrases pourra être examiné avec une rigueur qui, à l'heure actuelle, fait complètement défaut.

2.4.2. Pour vérifier l'efficacité d'un tel traitement, je l'ai appliqué au texte que j'avais utilisé pour mon exposé de Louvain-la-Neuve : César, *B. G.*, I, chap. 1 à 5. Le tableau suivant indique les longueurs des phrases dans l'ordre de succession du texte : dans la 1^{e} colonne, les chiffres correspondant à la découpe dépendancielle; dans la 2^{e} et la 3^{e} , ceux que l'on obtient en tenant compte de [.!?] puis de [.!? :;] dans l'édition de O. Seel (Teubner, 1968); dans la 4^{e} et la 5^{e} , ceux que procure l'édition de L. Constans (CUF, 1947) si on la traite de ces deux manières. J'ai marqué les correspondances en alignant horizontalement les données homologues. Pour faciliter la comparaison, j'ai uniformisé les petites différences textuelles : *iusiurandum* traité partout comme un seul mot, de même que *nihilominus*; en V, 4, omission de *suis* après *finitimis*, suivant l'édition de Seel.

[19] Sur ce point, voir mon étude, à paraître sous le titre "Pour un inventaire raisonné de la syntaxe latine", dans le volume que publiera le département d'études anciennes de l'Université de Liège à l'occasion du 175^{e} anniversaire de cette institution (1992).

Chap.	Synt. dép.	Seel [.!?]	Seel [.!?:;]	Constans [.!?]	Constans [.!?:;]
I	21	21	21	21	21
	8	8	8	8	8
	11	11	11	11	11
	71	42	42	42	42
		29	29	29	29
	13	30	30	30	30
	6				
	8				
	3				
	6	18	18	18	18
	6				
	6				
	16	22	22	22	16
	6				6
II	8	8	8	8	8
	33	33	23	33	23
			10		10
	49	49	11	49	11
			38		14
					12
					12
	24	24	15	24	15
			9		9
	27	27	27	27	27
III	38	38	38	38	38
	9	15	15	15	9
	6				6
	6	6	6	6	6
	6	6	6	6	6
	63	63	34	63	34
			29		29
	24	34	15	34	15
			9		9
	10		10		10
	24	24	24	24	24
IV	7	7	7	7	7
	8	15	8	15	8
	7		7		7
	28	35	28	35	28
	7		7		7
	32	32	21	32	21
			11		11
V	15	15	15	15	15
	23	23	23	52	52
	20	29	20		
	9		9		
	40	40	40	40	40

2.4.3. Les 704 mots du texte forment 38 phrases selon la découpe dépendancielle; la ponctuation de l'édition Seel donne 28 phrases dans l'un des systèmes et 38 dans l'autre; pour l'édition de Constans, les chiffres sont respectivement 27 et 40. Il est possible, à partir de ces données, de calculer la longueur moyenne de la phrase selon chaque système : respectivement 18,53; 25,14; 18,53; 25,89; 17,60.

Dans le cas particulier du texte ici pris en considération, c'est le système [.!? :;] appliqué à la ponctuation de Seel qui se rapproche le plus de la découpe dépendancielle, au point d'aboutir exactement au même nombre de phrases. Mais il serait téméraire, comme je l'expliquerai plus bas, de tirer de cette identiré des conclusions générales sur l'équivalence des deux méthodes.

Les cas où la découpe dépendancielle produit des phrases plus courtes que les autres systèmes sont au nombre de 8. Ce sont les passages où de simples virgules séparent des phrases indépendantes. Quant aux 7 cas où la découpe dépendancielle groupe des éléments que l'un au moins des autres systèmes disjoint, la divergence est due soit à la présence d'un relatif traité en relatif de liaison, soit à l'usage du point-virgule ou du double point pour distinguer des groupes non indépendants.

Comme on vient de le voir, la découpe dépendancielle et celle qui suit la ponctuation de Seel, en tenant compte des points-virgules et des doubles points, arrivent globalement au même résultat : 38 phrases. Cela ne signifie nullement qu'elles soient équivalentes : en fait, 7 phrases distinguées par la découpe dépendancielle se trouvent intégrées par Seel dans des phrases plus longues, tandis que, dans 7 autres endroits, des phrases de la découpe dépendancielle sont fragmentées par Seel, si bien que 7 phrases supplémentaires apparaissent. L'identité du résultat global est donc purement accidentelle.

2.4.4. Concluons sur ce point : la découpe dépendancielle devrait mettre fin à la controverse, puisqu'elle propose un critère rigoureux et scientifiquement fondé. Sans doute n'est-elle facilement praticable que pour les chercheurs qui disposent de fichiers informatisés comportant le descripteur de dépendance. Ce devrait être une incitation à multiplier les fichiers de ce type.

*

* *

3.0. Quant au traitement statistique des données, nous avions calculé la longueur moyenne et l'écart-type des phrases. M. Janson le signale sans commentaire. Il me paraît pourtant que l'on pourrait en faire un fort utile. G. U. Yule avait déjà observé que les distributions de longueurs de phrases sont dissymétriques et qu'elles

appartiennent à l'espèce où le carré de l'écart-type dépasse largement la moyenne[20]. Cela n'enlève pas à l'écart-type son caractère d'indice de dispersion, mais le fait qu'il s'agit de distributions dissymétriques le prive de la capacité prédictive qui est la sienne quand il concerne des distributions approximativement normales : on sait en effet que, dans de telles distributions, l'intervalle obtenu en reportant 1 écart-type de part et d'autre de la moyenne contient 68 % de l'effectif; si, au lieu de reporter 1 écart-type, on en reporte 2, c'est 95 % de l'effectif qui sont compris entre les limites ainsi définies. Pour les distributions de longueurs de phrases, la dissymétrie par rapport à la moyenne enlève toute signification à la propriété que je viens de rappeler.

3.1. Dans un article de 1939, C. B. Williams confirme les observations de Yule en étudiant d'autres échantillons que ce dernier, mais il note que, si on exprime la longueur de phrase non par le nombre de mots mais par le logarithme de ce nombre, on trouve une distribution qui ressemble assez à une distribution normale[21]. Cette particularité, si elle se confirme, est intéressante, entre autres motifs, parce que la nature même de l'échelle logarithmique a pour effet qu'un écart identique en nombre de mots a des valeurs logarithmiques différentes selon le niveau où elle se situe : soient deux phrases comptant respectivement 10 et 20 mots, les valeurs correspondantes en logarithmes décimaux sont 1.00000 et 1.30103; mais s'il s'agit de phrases de 100 et de 110 mots, on a 2.00000 et 2.04139. La transformation logarithmique réduit l'importance d'un même écart à mesure que l'on s'élève dans l'échelle des valeurs naturelles; c'est, me semble-t-il, de la même manière que varie l'effet produit sur le lecteur ou l'auditeur par la longueur des phrases et l'on aurait ainsi un moyen de traduire quantitativement cet effet[22]. J'ai donc testé la transformation logarithmique sur l'échantillon dont les données se trouvent plus haut. La moyenne logarithmique (calculée en log. décimaux) est 1.12743, ce qui, en valeurs naturelles, correspond à une longueur moyenne d'un peu plus de 13 mots. L'écart-type est 0.34368. Reporté de part et d'autre de la moyenne, il définit un intervalle dont les limites, traduites en nombres de mots, sont respectivement un peu supérieures à 6 et à 29. Cet intervalle contient 22 des 38 phrases, c'est-à-dire

---

[20] G. U. YULE, "On Sentence-length as a Statistical Characteristic of Style in Prose", dans *Biometrika*, 30 1938–1939, p. 363–390, spéc. 371.

[21] C. B. WILLIAMS, "A Note on the Statistical Analysis of Sentence-length as a Criterion of Literary Style", dans *Biometrika*, 31, 1939–1940, p. 356–361. Dans notre index de la *Consolation à Helvia*, p. 258, note, nous avions signalé l'article de Williams et manifesté notre intention de vérifier sur Sénèque la validité de la transformation logarithmique qu'il propose. En fait, dans le volume consacré à la *Consolation à Marcia*, nous avons utilisé du papier semi-logarithmique pour représenter la succession des longueurs de phrases, mais sans examiner les enseignements de ce mode de représentation.

[22] Cfr, à ce propos, les réflexions de L. DELATTE, *l.l.*, p. 60.

55 % du total. Le report de 2 écarts-types donne des limites un peu supérieures à 2 et à 65 mots, ce qui inclut 37 des 38 phrases, ou 97 % de l'effectif. Ces résultats, surtout si l'on tient compte de l'exiguïté de l'échantillon, sont fort proches de ceux que donnerait une distribution strictement normale. Il me paraît donc que la transformation logarithmique est probablement un instrument utile et fiable; il y aurait lieu de l'expérimenter plus largement que cela n'a été fait jusqu'ici. Je ne tenterai cependant pas de le tester sur les distributions des *Consolations* de Sénèque : le critère qui y est utilisé pour distinguer les phrases est, on l'a vu, peu satisfaisant; dès lors, il me paraît préférable d'attendre le moment où l'on disposera, pour ces œuvres, d'une découpe dépendancielle.

3.2.0. M. Janson consacre près de la moitié de son article à l'examen de notre étude des longueurs de phrases dans leur ordre de succession réelle. Pour obtenir une première vue, toute provisoire, des caractéristiques de cet ordre de succession, nous y avions appliqué la technique des moyennes mobiles. Comme on le sait, cette technique consiste à former, dans un ensemble, les sous-ensembles successifs de dimension x librement choisie, à calculer les moyennes respectives de ces sous-ensembles et à les rapporter chaque fois à leur phrase centrale. Nous avions choisi comme longueur 27 phrases; les sous-ensembles successifs étaient donc formés des phrases 1 à 27, 2 à 28, 3 à 29 etc, les moyennes étant rapportées successivement aux phrases 14, 15, 16 etc.

3.2.1. Comme on peut s'y attendre, tandis que le graphique des longueurs individuelles a une allure en dents de scie, celui des moyennes mobiles successives est beaucoup plus plat. Tel est le principal intérêt des moyennes mobiles : elles atténuent la diversité des données de base et font percevoir les tendances générales. C'est ainsi que nous avions pu distinguer, tant dans la *Consolation à Polybe* que dans la *Consolation à Helvia*, des variations portant sur des portions de texte qui se distinguent par leur thématique[23]. De ces observations, M. Janson ne souffle mot.

3.2.2. L'examen des moyennes mobiles nous apporta une autre observation, à laquelle nous ne nous attendions pas. Le voisinage immédiat des phrases les plus longues semble formé de phrases extrêmement courtes, qui, en quelque sorte, rétablissent l'équilibre. M. Janson, qui, dans notre recherche statistique, n'examine que la partie consacrée à ce sujet, considère que la longueur de 27 n'est pas appropriée à une telle étude[24] : à première vue, il paraît sans doute peu vraisemblable qu'une phrase, en raison de sa longueur, puisse exercer son action sur une portion de texte si importante. Mais, à l'inverse, on admettra que, dans un ensemble très

[23] Cfr SÉNÈQUE, *Consolation à Helvia - Index verborum - Relevés statistiques* (cité n. 1), p. 260 s.
[24] T. JANSON, *l.l.*, p. 33.

bref - disons de 5 phrases - la présence de phrases très courtes ne pourrait équilibrer pleinement une phrase très longue. De toute manière, nous avions fait un examen attentif des cas individuels[25], mais cet examen devrait être approfondi, par exemple par l'étude des phrases qui, dans les ensembles de 27, empêchent la moyenne d'augmenter malgré la présence d'une phrase particulièrement longue.

3.2.3. Il conviendrait d'ailleurs de reprendre la question à la lumière de ce qui a été dit plus haut. Tout d'abord, il serait désirable de ne traiter que des distributions établies selon la découpe dépendancielle : il semble qu'un travail sur des données d'un autre type serait vain. Par ailleurs, la transformation logarithmique proposée par Williams fournit certainement une base plus appropriée à la recherche que la longueur simplement évaluée en nombre de mots. Enfin, il serait indispensable de définir avec précision ce qu'est une phrase longue. Ce serait, par exemple, toute phrase dont la longueur, comptée en logarithme du nombre de mots, vaudrait au moins la moyenne plus $1^1/_2$ écart-type. Dans une distribution normale, cela correspond à 7 % de l'effectif total, ce qui me paraît une bonne proportion. Il resterait alors à examiner la composition du groupe dont chaque phrase longue serait le centre.

3.2.4. En lieu et place du calcul de la moyenne mobile, M. Janson propose une méthode qui, en revanche, ne me paraît nullement adéquate : il suggère l'emploi du calcul des "séquences" ou "suites" (*runs*). Si l'on répartit les éléments d'un ensemble en deux sous-ensembles strictement définis et que l'on observe l'ordre dans lequel ces éléments apparaissent, on appelle *run* ("séquence" ou "suite") le groupe d'éléments en succession immédiate qui appartiennent tous au même sous-ensemble. Si l'on connaît le nombre d'éléments de chacune des deux catégories, on peut déterminer dans quelle mesure le nombre de séquences est supérieur ou inférieur à ce que produirait le simple hasard[26]. Les raisons pour lesquelles cette méthode ne me paraît pas susceptible d'être appliquée à la question de l'environnement des phrases longues sont au nombre de trois. En premier lieu, comment répartir les phrases en deux types? M. Janson ne le dit pas explicitement, mais les chiffres qu'il cite donnent à penser qu'il a cherché à obtenir deux effectifs aussi proches que possible l'un de l'autre[27]. C'est là une répartition tout à fait artificielle : elle ne met pas en lumière un caractère qui opposerait nettement deux types de phrase[28].

---

[25] Cfr *Consolation à Helvia* (cité n. 1), p. 262 s.

[26] T. JANSON, *l.l.*, p. 34, qui, pour la théorie, renvoie au traité de P. G. HOEL, *Elementary Statistics*, 1960. Pour sa part, L. DELATTE, *l.l.*, p. 57, s'appuie sur S. SIEGEL, *Nonparametric Statistics for the behavioral Sciences*, 1956. Quant à moi, je citerais volontiers l'exposé, en français, de J.-M. FAVERGE, *Méthodes statistiques en psychologie appliquée*, t. 2 (1972), p. 323 s.

[27] T. JANSON, *l.l.*, p. 35, n. 1. Cfr L. DELATTE, *l.l.*, p. 58.

[28] En fait, dans la plupart des textes, il y a lieu de distinguer trois groupes à phrases : les phrases remarquables par leur longueur; celles qui le sont par leur brièveté; celles dont la longueur n'a rien de

Dès lors, les séquences délimitées de cette façon n'ont guère de sens. En second lieu, les calculs relatifs aux séquences concernent la totalité de l'effectif, alors que nous avions porté notre attention exclusivement sur les phrases particulièrement longues et sur leur environnement immédiat : pour la *Consolation à Polybe*, par exemple, des 235 phrases, 42 avaient été retenues comme particulièrement longues et l'étude de leur environnement n'impliquait que 143 phrases, c'est-à-dire 61 % du total; il n'y a donc pas adéquation dans les données. En troisième lieu, s'il est vrai qu'une longue phrase entourée de phrases courtes constitue presque certainement une séquence à elle seule (quel que soit le critère utilisé pour distinguer les deux types de phrase), cela ne donne aucune indication sur les effectifs des séquences de phrases brèves qui l'entourent; ces séquences pourraient être fort courtes ou se prolonger sur un grand nombre de phrases. En somme, la méthode recommandée par M. Janson conviendrait plutôt à l'étude de notre problème précédent : examiner dans quelle mesure les diverses parties de l'œuvre se diversifient quant à la longueur des phrases. Encore faudrait-il avoir trouvé un critère acceptable de répartition de l'effectif en deux classes[29].

*

* *

4. Il me reste à conclure. En premier lieu, je l'ai déjà dit, la syntaxe dépendancielle fournit le moyen de délimiter les phrases en fonction de critères assurés. Par ailleurs, la transformation logarithmique paraît correspondre à la nature des données mieux que ne le fait l'usage des valeurs naturelles. Enfin, pour ce qui concerne l'environnement des phrases particulièrement longues, les premières indications que nous avions recueillies devraient être complétées par des examens plus précis, mais la méthode proposée à cette fin par M. Janson n'est pas adéquate.

remarquable. En distinguer deux, de quelque manière que l'on s'y prenne, reviendrait à créer au moins un et probablement deux sous-ensembles composites, dont les caractères distinctifs n'auraient aucune portée essentielle; c'est une situation toute différente de celle que l'on aurait si, par exemple, on était amené à distinguer, dans un ensemble de pièces d'un échiquier, les blanches et les noires, ou encore, dans une file d'attente, les femmes et les hommes.

[29] L'article de J.M. MICHEL, "Premiers éléments d'une stylistique quantitative : la longueur de la phrase dans l'*Agricola* de Tacite" dans *Initiation à la linguistique contemporaine*, t. 3, 1982, 103–127, repris dans J.H. MICHEL, *Le latin aujourd'hui*, Bruxelles, 1985/6, p. 76–87, surtout p. 79, s'intéresse au problème des longueurs de phrases, mais dans un sens différent.