Comment mesurer l'étendue d'un texte?

Autor(en): **Muller, Charles**

Objekttyp: **Article**

Zeitschrift: **Revue de linguistique romane**

Band (Jahr): **33 (1969)**

Heft 131-132

PDF erstellt am: **22.09.2024**

Persistenter Link: https://doi.org/10.5169/seals-399448

COMMENT MESURER L'ÉTENDUE D'UN TEXTE?

Quiconque s'est intéressé à la statistique lexicale sait que l'une des faiblesses de cette discipline, celle que l'on mentionne le plus souvent, réside dans la difficulté de définir de façon satisfaisante les unités de base de toutes ses opérations : le mot, unité de texte et le vocable, unité de lexique; on ne se tire d'affaire et on n'obtient des résultats acceptables qu'au prix de conventions précises dont l'ensemble constitue une norme lexicale [1].

Mais il est une autre source d'incertitude et d'inquiétude, généralement passée sous silence, et qui atteint les applications de la statistique non seulement au lexique, mais à n'importe quel objet linguistique : c'est la mesure de l'étendue des textes. En effet, la fréquence d'un fait de langue quelconque (phonétique, morphosyntaxique, lexical, stylistique,...) n'a de signification que si elle est rapportée à l'étendue du texte considéré. Dire qu'un fait est plus fréquent dans le texte A que dans le texte B ne saurait s'appliquer à la fréquence absolue, mais à une fréquence relative qui ne peut être établie que si l'on connaît avec une précision suffisante le rapport de longueur des deux textes.

Dans bien des cas, certes, on peut se contenter d'une estimation, d'une approximation : on compte des pages ou des lignes, en tenant compte tant bien que mal des différences typographiques, et on arrive à une mesure rudimentaire : A est deux fois et demie plus long que B, par exemple, ou quelque chose de semblable; et c'est souvent bien suffisant.

Mais la rigueur et le raffinement de certaines opérations statistiques, et en particulier des travaux qui prennent pour objet des fragments très courts, exigent une précision de plus en plus grande, comme cela s'est passé dans d'autres domaines scientifiques, et nous conduisent à mettre

1. Voir à ce sujet « Le mot, unité de texte et unité de lexique en statistiqne lexicologique », dans *Travaux de Ling. et de Litt.*, 1963, et *Initiation à la statistique linguistique*, Larousse, 1968, p. 142 à 151.

en cause les moyens de mesure utilisés pour déclarer que « A est *n* fois plus long que B ».

On peut soit considérer le texte dans sa totalité, soit se limiter à certains de ses éléments pris comme base. Ainsi quand on veut mesurer la fréquence relative de la voyelle *a*, on peut rapporter sa fréquence absolue soit au nombre des phonèmes du texte, donc à la totalité de ce texte, soit seulement au nombre des voyelles, voire des voyelles orales. La fréquence d'un verbe peut s'évaluer soit par rapport au nombre des mots du texte, soit en prenant comme base les seules formes verbales. Ainsi, constatant que le verbe *faire* a 137 formes dans le *Cid* et 52 dans *Phèdre*, je puis mettre ces fréquences absolues en relation avec les 3 409 formes verbales relevées dans la première de ces pièces et les 3 005 de l'autre [1] ; d'où des fréquences relatives de :

$$137/3\,409 = 0{,}040$$
$$52/3\,005 = 0{,}017$$

qui deviennent comparables entre elles, et qui permettent d'avancer que ce verbe est 2,35 fois plus fréquent dans la langue de la première pièce que dans celle de l'autre. Mais en tenant compte du nombre total d'occurrences qui constituent ces deux textes, je trouverais :

$$137/16\,424 = 0{,}00834$$
$$52/14\,217 = 0{,}00366$$

soit un rapport de 2,28, qui diffère d'ailleurs fort peu du précédent; ce qui permet de constater que dans les deux textes la densité en formes verbales varie peu (0,208 dans le *Cid*, 0,211 dans *Phèdre*), mais que le verbe *faire* a des fréquences nettement différentes [2].

Si l'on n'avait pas eu les données utilisées ci-dessus, on aurait pu aussi se référer au nombre de vers, et l'on aurait trouvé que dans le *Cid* notre verbe apparaît en moyenne 7,23 par 100 vers, dans *Phèdre* 3,14 fois, soit un rapport de 2,30, encore très voisin des résultats précédents.

Quand on décide de mesurer le texte dans sa totalité, sans critère

1. Ces données sont empruntées aux *Concordances* de ces pièces (Larousse, 1966).

2. Par 100 vers, dans l'ensemble du théâtre de Corneille, 8,56; dans ses pièces classées en trois périodes chronologiques : 7,77, 8,81, 8,90; dans ses tragédies : 8,86; dans ses comédies : 8,59. Chez Racine : *Bajazet*, 4,00; *Britannicus*, 3,84; *Bérénice*, 4,38; *Andromaque*, 5,65, etc.

sélectif, et quand on opère sur un texte imprimé [1] (en général indexé), on peut prendre pour base :

— soit des unités graphiques, comme les lettres, les lignes, les pages;
— soit des unités phoniques, comme les phonèmes, les syllabes, les vers;
— soit des unités sémantiques, comme les mots, les propositions, les phrases.

Les unités graphiques sont aisément décomptées quand le texte est traité en mécanographie, par un comptage mécanique des mots de 1, de 2, ... de *n* lettres; quant au compte des lignes ou des pages, il se ramène en fin de compte à un dénombrement de lettres et d'espaces, auxquels s'ajoutent les ponctuations; car si l'on doit comparer des textes de typographie différente, il faudra bien évaluer, dans chacune des impressions en cause, le nombre moyen de signes par ligne, et compter les lignes par page; l'estimation est faite alors par un sondage aléatoire, assorti si possible d'un calcul d'erreur type.

Il est rare que l'on puisse prendre pour unité le phonème, à moins de réaliser une transcription phonétique du texte [2]; en français surtout, les désaccords entre la longueur graphique et la longueur phonique des mots et groupes de mots sont grands, et des expériences seraient utiles pour en connaître les incidences; elles manquent actuellement. En revanche, la syllabe sert très facilement de base quand on compte par vers, à condition de tenir compte de la longueur de chaque vers. Ainsi le *Cid* a 1 840 vers, ce que toute édition classique nous apprend immédiatement; mais les stances de Rodrigue (acte I, sc. 6) et celles de l'Infante (V, 2) introduisent parmi les alexandrins 18 décasyllabes, 18 octosyllabes et 12 vers de 6 syllabes, que nous compterons respectivement pour 15, 12 et 6 alexandrins; soit, pour ces 48 vers plus courts, une équivalence de

1. On ne mentionne que pour mémoire la possibilité de mesurer par sa durée un texte écouté ou enregistré, cette durée étant fonction de la façon de dire autant que du contenu. La presse nous a habitués, depuis quelques années, à voir évaluer la durée d'un discours politique, d'une déclaration importante en nombre de mots plutôt qu'en minutes.

2. C'est ce qu'ont fait, pour plusieurs langues, M. Guiter et ses collaborateurs, mais pour des lemmes et non des textes (« Corrélations de signifiants et de signifiés », *Trav. de Ling. et de Litt.*, VII, 1, 1969, p. 131-159).

33 alexandrins; on comptera donc pour cette pièce 1 840 + 33 — 48 = 1 825 alexandrins. Cette correction est aisée dans des œuvres où la très grande majorité des vers est isomètre; elle devient laborieuse pour les pièces classiques en vers libres (*Agésilas*, *Amphitryon*...); elle est quasi impossible dans des textes modernes comme ceux d'Apollinaire.

L'unité qui sert le plus couramment à mesurer l'étendue d'un texte est le mot, unité sémantique, et il est déjà traditionnel de représenter par N le nombre des occurrences qui constituent le texte (V étant le nombre de vocables ou mots différents).

La question est évidemment de savoir s'il est indifférent de choisir (ou de se laisser imposer par les circonstances) l'un ou l'autre de ces procédés. Ils seraient équivalents si, dans les textes sur lesquels on opère, la longueur moyenne du mot en syllabes était constante, et s'il en était de même de la longueur graphique moyenne du mot. Mais l'expérience prouve qu'il n'en est pas nécessairement ainsi, et que ces moyennes sont en fait des indices stylistiques importants, des variables dont l'effet peut devenir sensible. J'ai montré ailleurs que chez Corneille [1], par exemple, le nombre moyen de mots par vers (donc par 12 syllabes) n'a cessé d'augmenter avec l'âge de l'auteur : légèrement inférieur à 9 dans les œuvres de jeunesse (*Médée* : 8,800), il atteint 9,530 dans sa dernière pièce ; ce qui revient à dire que le nombre moyen de syllabes par mot a passé de 1,36 à 1,26 : l'écart entre tragédies et comédies, pour cette même moyenne, est moins frappant, mais non moins certain.

M. Pelchat, au cours de son étude statistique sur les rôles de l'*Avare* [2], a tenté de comparer la mesure fournie par le nombre de mots, N, avec celle qu'il obtient avec le nombre de lettres, L ; la comparaison des deux valeurs donne un nombre moyen de lettres par mot qui, dans cette pièce, est de 3,939 ; dans les rôles, si l'on écarte les plus courts (moins de 500 mots), cette moyenne varie entre 3,713 (monologue d'Harpagon) et 4,135 (rôle d'Anselme). Il est probable que ces variations ont une signification stylistique, en rapport avec la densité du texte en mots de relation. Mais on voit aussi que quand il s'agit d'évaluer l'étendue d'un des rôles par rapport à celle de la pièce, ce qui est ici notre objet, le résultat varie peu suivant que l'on a pris pour base les mots ou les lettres :

1. V. *Étude de statistique lexicale*, Larousse, 1967, p. 49-52.

2. Roland Pelchat, *L'Avare de Molière, étude statistique du vocabulaire des rôles*, Strasbourg, 1969 (thèse dactylographiée).

	N'/N (mots)	L'/L (lettres)	écart abs.	écart rel.
Harpagon (dial.). . .	0,2637	0,2619	0,0018	0,0067
Cléante.	0,1640	0,1655	0,0015	0,0091
Valère.	0,1350	0,1355	0,0005	0,0037

etc. Mais avec les fragments où l'écart entre les deux mesures est le plus fort, on obtient :

Harpagon (monol.) .	0,0368	0,0347	0,0021	0,0605
Anselme.	0,0252	0,0264	0,0012	0,0455

ce qui fait que si l'on veut évaluer la longueur de ces deux fragments l'un par rapport à l'autre, on arrive à un rapport de 1,46 en comptant les mots, de 1,31 en comptant les lettres : écart appréciable dû à une différence stylistique très nette [1]. Dans des cas de ce genre le choix de la base est décisif. Le monologue d'Harpagon contient 100 substantifs, le rôle d'Anselme 72 : suivant la base adoptée, on dira que c'est l'une de ces fréquences qui l'emporte, ou l'autre.

Voici maintenant une expérience qui vise à déterminer si la mesure fondée sur les vers, donc sur une base syllabique, doit être préférée à celle qui prend le mot comme unité.

Ayant travaillé sur deux textes indexés, deux tragédies de la vieillesse de Corneille, *Sertorius* et *Sophonisbe*, je dispose de 37 tranches de 100 alexandrins chacune (19 de *Sertorius*, 18 de *Sophonisbe*), et je connais pour chacune d'elles N, le nombre de mots, et V, le nombre de vocables qu'elle contient.

1. J'ai tenté d'ajouter à ces données un comptage de syllabes. Le résultat est que le nombre de lettres par syllabe, dans les deux fragments, est sensiblement le même : 3,243 et 3,230. Ce qui diffère entre eux, et qui par conséquent peut servir d'indice stylistique, c'est la longueur moyenne du mot, tant phonique (1,144 et 1,279 syllabes) que graphique (3,713 et 4,135 lettres). Il en résulte que si l'on se fonde sur l'étendue phonique (nombre de syllabes) ou graphique (nombre de lettres), on trouve que le premier est 1,310 ou 1,304 fois plus long que l'autre, ce qui est sensiblement égal ; mais si l'on se fonde sur le nombre de mots, le rapport passe à 1,459, ce qui n'est plus du tout la même chose. De même, si l'on emploie soit les syllabes, soit les mots pour comparer l'étendue de *Sertorius* et *Sophonisbe*, pièces proches par leurs dates, leurs sujets et leur genre, on obtient 1,048 ou 1,055, ce qui diffère peu ; mais si l'on rapproche *Médée* et *Suréna* (39 ans d'intervalle), on chiffre le rapport de leurs étendues respectives soit à 1,073, soit à 1,162, donc avec un écart de 8 % entre les deux estimations.

La valeur N apparaît comme une variable approximativement gaussienne, de moyenne $\overline{N} = 925{,}84$ et d'écart type $\sigma = 19{,}46$:

N	nombre de tranches
880-889	1
890-899	3
900-909	3
910-919	7
920-929	10
930-939	4
940-949	5
950-959	2
960-969	1
970-979	1
	37

Si l'on traite séparément les deux pièces, on trouve pour *Sertorius* $\overline{N} = 922{,}68$ et $\sigma = 18{,}12$; pour *Sophonisbe* 929,17 et 20,26. Les valeurs extrêmes sont 888 mots pour les vers 1201 à 1300 de *Sophonisbe* (891 pour les vers 1801 à 1900 de *Sertorius*), et 977 pour les vers 1201 à 1300 de *Sertorius* (963 pour les vers 1603 à 1702 de *Sophonisbe* [1]). Dans ces cas extrêmes, des fragments considérés comme d'égale étendue d'après le critère syllabique auraient entre eux des différences d'étendue atteignant 10 % en prenant le critère lexical.

Dans ces tranches, l'étendue du vocabulaire, V, nombre de mots différents, est une variable un peu plus dissymétrique que N, de moyenne $\overline{V} = 314{,}14$, et d'écart type $\sigma = 15{,}40$:

V	nombre de tranches
280-289	1
290-299	4
300-309	13
310-319	6
320-329	6
330-339	5
340-349	1
350-359	1
	37

1. La tranche précédente, comprenant 5 vers courts, est prolongée de 2 alexandrins, pour être égale aux autres, et s'arrête au vers 1602 ; la pièce qui compte 1 822 vers, est comptée pour 1820 alexandrins.

Dans les deux pièces, on trouve 315,21 et 17,50 pour *Sertorius*, 313,00 et 12,70 pour *Sophonisbe*. Valeurs extrêmes : 287 (v. 1001 à 1100) et 359 (v. 1 à 100) pour l'une, 290 (v. 601 à 700) et 334 (v. 401 à 500) pour l'autre.

On recommence les mêmes opérations sur V en prenant cette fois des tranches où N, nombre de mots, sera constant ; pour cela, reprenant les 37 tranches précédemment définies, on les ajustera toutes à 925 mots, en retranchant leurs derniers mots ou en ajoutant les mots qui suivent immédiatement, suivant que 925 — N est positif ou négatif. On observe les modifications de V et on compare les résultats nouveaux avec ceux qui ont été obtenus précédemment :

	base syllabique		base lexicale	
	$\overline{V}$	σ	$\overline{V}$	σ
Sertorius....	315,21	17,50	315,47	19,29
Sophonisbe..	313,00	12,70	311,83	12,87
ensemble....	314,14	15,40	313,70	16,59

Les variations sur la moyenne sont normales. En effet, dans la première pièce, où la moyenne était inférieure à 925, on a augmenté le texte de 44 mots pour aligner les tranches à cette valeur ; d'où une légère hausse de la moyenne de V ; dans l'autre, où la moyenne de N était supérieure à 925, on a retranché au total 75 mots, ce qui explique la baisse du V moyen ; pour l'ensemble, c'est donc une diminution de 31 mots qui s'est répercutée en faisant légèrement baisser la moyenne de V.

Mais l'intérêt est dans la hausse des écarts-types, donc de la dispersion des valeurs de V, et cela dans les deux textes et, évidemment, dans l'ensemble qu'ils forment.

On pourrait donc avancer que, dans ce cas au moins, le contenu lexical s'est révélé plus stable dans les tranches de 100 vers que dans celles de 925 mots, et qu'en conséquence l'unité syllabique est au moins aussi valable que l'unité lexicale.

Le phénomène est du reste facile à expliquer. Dans un cadre syllabique fixe, les variations de N sont déterminées surtout par la densité plus ou moins forte en mots courts, donc en mots de relation. Si bien que la tranche qui contient le plus de mots est loin d'être celle qui contient le plus de vocables. En isolant une trentaine de vocables qui arrivent en tête dans la liste des fréquences, qui représentent exactement 50 % du

texte et qui sont presque exclusivement monosyllabiques [1], j'ai trouvé entre leur fréquence (tant absolue que relative) et V une corrélation inverse très forte, de l'ordre de — 0,70.

Certains travaux statistiques ne considèrent que les « mots forts », à l'exclusion des mots-outils ou mots de relation; mais la distinction entre ces deux catégories n'est pas toujours claire. Le procédé a cependant l'avantage de donner probablement une image plus juste du contenu réel du texte que celui qui accorde une égale valeur à toutes les occurrences.

La mesure syllabique, donnant à chaque occurrence le poids du nombre de ses syllabes, se rapproche peut-être mieux encore d'une juste appréciation de l'étendue du texte.

L'idéal serait sans doute de disposer toujours, en vue de ces différentes mesures, des bases graphique, phonique et lexicale qui viennent d'être examinées : leurs désaccords éventuels, à l'intérieur d'un même idiome, donneraient souvent une indication stylistique utile; et il serait intéressant, en soumettant à cette épreuve la traduction d'un même texte en plusieurs langues, de déterminer si les rapports entre ces trois données peuvent caractériser, distinguer et classer les idiomes, comme d'autres critères quantitatifs ont déjà permis de le faire avec succès [2].

Mais dans la pratique, il en est rarement ainsi. Suivant la nature du texte, suivant son étendue, suivant enfin les moyens disponibles (comptage « à la main », consultation d'un index ou d'une concordance, traitement mécanographique de cartes perforées, etc.), c'est l'une ou l'autre de ces données qui sera accessible. Du moins est-il bon de ne pas les accepter indifféremment, et de s'interroger sur leur valeur réelle.

Ces incertitudes et ces options ne sont pas propres aux applications

1. Ils comprennent des mots toujours monosyllabiques (*et*, *en*, *mon* et sa flexion, etc.); des mots à élision, comptant tantôt pour 1 syllabe, tantôt pour 0 (*le*, *de*, *ne*, *que*, etc.); enfin quelques verbes (*être*, *avoir*, *faire*...) qui ont des formes polysyllabiques, mais dont les formes les plus fréquentes sont monosyllabiques. A noter que si dans ces deux pièces on retient les vocables dont la fréquence dépasse 100, on trouve le même nombre et, à une unité près, les mêmes vocables.

2. V. surtout les travaux de G. Herdan, et les articles récents de M. H. Guiter : « Quelques paramètres caractéristiques des systèmes vocaliques » (*RLiR*, janv.-juin 1966) et « Concordances linguistiques et anthropologiques » (*ibid.*, janv.-juin 1969), ainsi que l'article cité ci-dessus, p. 247, n. 2.

linguistiques de la statistique [1] ; dès que cette méthode est appliquée non à des artifices comme les jeux de hasard (pile ou face, dés, cartes...) mais aux réalités de la nature et de la vie humaine, elle soulève des questions de norme ou de base; si l'on dit qu'il y a plus d'accidents d'auto dans le pays A que dans le pays B, cela suppose d'abord que l'on a accepté une définition de l'unité « accident » (comme nous devons, tant bien que mal, définir nos unités « mot », « syllabe », etc.); ensuite que l'on a choisi une base, car il ne s'agit évidemment point de comparer le nombre absolu des accidents en A et en B, mais de rapporter ces nombres à une mesure de l'importance respective des deux pays. Prendra-t-on leur superficie, leur population, l'effectif de leur parc automobile, la longueur (et la qualité ?) de leur réseau routier, la circulation appréciée d'après la consommation d'essence? ou bien une combinaison pondérée de ces diverses données? En tout cas, pas de comparaison, pas de conclusion sans une décision préalable sur ces problèmes. En linguistique comme ailleurs, il vaut mieux les aborder que les éluder.

Charles MULLER.

1. Cette conclusion reprend une note de mon *Initiation à la statistique linguistique*, p. 208.