L'entropie comme indice de diversification du vocabulaire dans les tragédies de Sophocle

Georges RIGO

Abstract. In order to characterize the richness of any work's vocabulary, the calculation of the entropy over increasingly longer text blocks provides an interesting index. The curve of resulting values can be adjusted to a theoretical parabolic function. The programme described enables the calculation of the entropy over any text as far as its forms are lemmatized. When applied to Sophocles' tragedies, this technique allows several interesting observations. Some indications are observed coinciding with chronological data, which distribute these works into two distinct groups; the variations in the resulting values of entropy are due mainly to significant words and not to functional words; they correspond to the work's subject matters and not to language registers.

Keywords: Lexicography, richness of vocabulary, diversity index, greek literature, Sophocles, content analysis, entropy, thematic vocabulary.

Mots-clés : Lexicographie, richesse du vocabulaire, indice de diversité lexicale, littérature grecque, Sophocle, analyse du contenu, entropie, vocabulaire thématique.

La diversité du vocabulaire d'un auteur peut être considérée comme un signe distinctif de son époque, voire de sa personnalité; cette diversité en effet est d'une certaine façon un signe de la « richesse » ou de la « pauvreté » de la langue dans laquelle il s'exprime, ou encore un caractère particulier du domaine culturel qui le préoccupe (texte poétique, philosophique, militaire...); nous pouvons aussi y voir l'effet d'une recherche personnelle pour se démarquer de ses contemporains, en raison du sujet traité, ou de la forme littéraire adoptée.

Dans le monde de l'Antiquité grecque, les V^e^–IV^e^ siècles av. J.-C. sont une période où, de l'avis général, la diversification volontaire du vocabulaire a connu la faveur des écrivains, notamment sous l'influence des Sophistes. Encore faut-il remarquer que c'est là aussi la période dite « classique », celle dont nous avons conservé un grand nombre d'œuvres de qualité dans tous les domaines

✉ Rue Justin Lenders, 89; B–4020 Liège (Belgique).

littéraires et où, par conséquent, nous connaissons le plus grand nombre de mots dans tous les domaines de la pensée.

Il existe bon nombre de techniques et de tests statistiques dont le but est de mettre en évidence cet aspect de la lexicographie. À diverses reprises, et notamment lors de la XI[e] Conférence Internationale sur l'Ordinateur et les Recherches Linguistiques[1], Étienne Évrard a suggéré d'utiliser à ce propos le calcul de l'entropie[2] en considérant successivement des fragments de plus en plus étendus de l'œuvre considérée.

En effet, le simple examen de la fréquence moyenne d'emploi de chaque lemme (nombre total d'occurrences rapporté au nombre total de vocables[3] constituant le texte), ne fournit aucune indication relative à l'évolution de cette diversité au fil de l'œuvre; en outre, la fréquence moyenne d'emploi verra nécessairement son évolution se ralentir en fonction de la longueur de l'œuvre, puisque le vocabulaire de la langue n'est pas infini.

Si on relève les *semel dicta* (en y ajoutant éventuellement les lemmes employés deux ou trois fois seulement), on considère certes un aspect important de la recherche de diversification, mais on ne tient toujours pas compte de l'endroit de l'œuvre où se situent ces mots (sans compter que des mots « diversifiants » peuvent être affectés d'une fréquence moyenne d'emploi assez élevée, parce que toutes les occurrences sont rassemblées dans une même partie de l'œuvre).

C'est pourquoi l'entropie, qui est le produit de deux quantités dont l'évolution obéit à des raisons mathématiques différentes (linéaire pour l'une, logarithmique pour l'autre) paraît plus apte à cerner les différents aspects de cette caractéristique lexicale.

Étienne Évrard a montré[4] comment la somme des entropies évolue selon que le mot examiné est une occurrence d'un lemme déjà rencontré, ou la première apparition d'un nouveau lemme.

---

[1] ÉVRARD (Ét.) : 1985, « Richesse et mode d'enrichissement d'un vocabulaire », *L'ordinateur et les recherches littéraires et linguistiques, Actes de la XI[e] Conférence Internationale — Louvain 1984* (Paris).

[2] Pour rappel, l'entropie se calcule en multipliant la probabilité (ici, la fréquence c'est-à-dire le rapport du nombre d'occurrences d'un lemme au nombre total d'occurrences dans le passage considéré) d'une quantité par le logarithme de cette probabilité : $H = -p_i \times \log p_i$. En thermodynamique, l'entropie exprime le degré de stabilité d'un état; en cybernétique, elle définit le degré d'incertitude d'un message en fonction des éléments transmis antérieurement; dans ce cas, quand l'entropie est nulle il n'y a pas d'incertitude.

[3] Le terme *vocable* (dans la suite, on emploiera de préférence le mot *lemme*, par référence au dictionnaire de Liddell et Scott), désigne une unité lexicale, tandis que *mot* ou *occurrence* désigne une unité lexicale employée dans un contexte.

[4] *L.c.*, p. 149.

En calculant l'entropie pour des fragments de plus en plus longs de l'œuvre, on obtient une suite de valeurs formant une courbe d'aspect parabolique, que l'on peut ajuster à une courbe théorique de formule $y = A x^2 + B x$.

On peut alors déterminer pour cette fonction la valeur des coefficients A et B, qui représentent respectivement la diversification (le coefficient B tend à éloigner la courbe de l'axe des x) et l'essoufflement (le coefficient A tend à rapprocher la courbe de l'axe des y) caractéristiques de l'œuvre[5].

Pour illustrer sa communication, Ét. Évrard présentait une série de valeurs obtenues en calculant ces facteurs pour différents textes d'auteurs latins (notamment César, Cicéron, Sénèque, Salluste). Le programme qu'il avait à l'époque écrit en PL/1 utilisait un ordinateur où la puissance s'alliait à une taille imposante.

Il souhaitait aussi voir sa proposition mise à l'épreuve d'autres domaines linguistiques. Or, l'essort récent des micro-ordinateurs, et la miniaturisation qui s'en est suivie, permettent aujourd'hui de traiter des fichiers importants — où sont enregistrées des œuvres complètes — sur des machines qui n'occupent qu'une partie de la surface d'une table de bureau. Il était donc intéressant d'écrire un programme réalisant sur un micro-ordinateur les différentes étapes nécessaires au calcul de l'entropie. C'est ce que j'ai réalisé en utilisant le langage Turbo Pascal 6.0.

Je souhaitais un outil aussi souple que possible, qui permettrait par exemple une utilisation indépendante de la langue des textes examinés. L'algorithme tire parti au maximum de la mémoire vive de l'ordinateur, ce qui permet notamment de ne pas définir au préalable les dimensions des tableaux de données résultant de l'examen du texte et accélère évidemment le traitement.

Les opérations assurées par le programme sont les suivantes :

— lecture du texte par pas successifs de 500 mots[6], avec classement des différents lemmes rencontrés et comptage de leur nombre d'occurrences;
— à la fin de chaque groupe, évaluation, pour chaque lemme, de sa probabilité en fonction du nombre total de mots déjà lus, et calcul de l'entropie correspondante;
— à chaque étape et à la fin du texte, calcul des coefficients A et B par approximation selon la méthode des moindres carrés;

---

[5] Le coefficient C de la formule générale de la parabole est ici égal à 0, puisque les valeurs partent toujours de zéro (dans le cas théorique d'un texte composé d'une seule occurrence, la probabilité du lemme serait égale à 1, et le logarithme égal à 0; la somme des entropies serait nulle).

[6] Cette valeur est ajustable selon les désirs de l'utilisateur.

— enregistrement des valeurs ainsi obtenues (qui s'affichent au fur et à mesure sur l'écran) dans deux fichiers ASCII destinés à être récupérés éventuellement par un tableur, un logiciel graphique ou un traitement de textes[7].
— enfin, si on le souhaite, le programme crée un fichier de lemmes qui représente le « lexique » de l'œuvre considérée : on y trouve la liste complète de tous les lemmes employés, avec pour chacun l'indication du nombre total d'occurrences, et le numéro d'ordre de la forme où il apparaît en premier lieu.

Ainsi réalisé, le programme permet de calculer les coefficients de n'importe quelle œuvre, pourvu que les enregistrements répondent à un minimum de critères de conformité :
— chaque mot doit faire l'objet d'un enregistrement se terminant par les caractères ⟨CR⟩ ⟨LF⟩ ;
— la position et la longueur du lemme doivent être constantes à l'intérieur de cet enregistrement, de même que la position d'un code de catégorie grammaticale[8].

*
* *

Le L.A.S.L.A. dispose du texte complet de Sophocle (sept tragédies, les fragments tragiques et les fragments lyriques), enregistré à partir du *Thesaurus Linguae Graecae*, et lemmatisé de manière semi-automatique par les programmes du Laboratoire, en vue de la préparation d'un index informatisé. J'ai donc appliqué ce programme à l'examen des sept tragédies de Sophocle, ce qui a permis quelques constatations intéressantes.

Diversité du vocabulaire et chronologie des œuvres

Le calcul par approximation (méthode des moindres carrés) aboutit à une somme des carrés des différences comprise entre 0,008 (*Philoctète*) et 0,062

[7] L'un est une copie ASCII de l'écran ; chaque ligne affichée constitue un enregistrement terminé par ⟨CR⟩ ⟨LF⟩. Le second ne reprend que les données chiffrées, séparées par des signes « ; », le séparateur décimal étant la virgule ; chaque ligne est terminée par ⟨CR⟩ ⟨LF⟩.

[8] Ce code, sous forme d'une lettre majuscule, sera ajouté au lemme pour distinguer les mots homonymes de catégorie grammaticale différente (p. ex. préposition / adverbe / conjonction). Les caractères reconnus sont les majuscules de A à Z, les chiffres de 1 à 9, et le caractère « espace » (ASCII 32).

(*Antigone*), ce qui indique un très bon ajustement de la parabole théorique à la courbe des valeurs observées. Les valeurs obtenues pour les coefficients A et B de l'équation

$$y = Ax^2 + Bx$$

sont comprises respectivement pour A entre $-0{,}047$ et $-0{,}064$ et pour B entre 1,12 et 1,25. Les valeurs extrêmes sont celles relevées pour *Ajax* (minima) et *Électre* (maxima).

Le classement des tragédies par ordre croissant de la valeur approchée de B (et de A, sauf pour les *Trachiniennes*), donne l'ordre repris au tableau 1.

Tableau 1

	A	B	Date
Antigone	0,04761	1,12634	442
Œdipe à Colone	0,05219	1,14514	?
Les Trachiniennes	0,05076	1,15688	?
Ajax	0,05442	1,18968	438
Œdipe Roi	0,06148	1,22264	420
Philoctète	0,06186	1,22817	409
Électre	0,06476	1,25799	415

Si l'on excepte *Œdipe à Colone*, dont on sait seulement que la date de représentation est postérieure à la mort de l'auteur, sans aucune indication sur la date de composition, et *Philoctète*, (dont l'analyse provoque l'étonnement des critiques quand ils rappellent l'âge de l'auteur à l'époque de la représentation), cet ordre correspond assez bien à la chronologie connue[9].

En reportant sur un même graphique les sept courbes obtenues par ajustement du calcul des entropies à partir des coefficients A et B, on voit apparaître nettement deux familles distinctes de courbes : la première regroupe *Antigone*, les *Trachiniennes*, *Œdipe à Colone* et *Ajax* (pour ces œuvres, B est

[9] La date de l'*Antigone* nous est connue par le témoignage d'Aristophane de Byzance; celle d'*Ajax* résulte d'une allusion vraisemblable aux *Crétoises* d'Euripide; *Œdipe Roi* est sans doute de peu postérieur à la peste d'Athènes (430); selon la didascalie de *Philoctète*, la tragédie a été représentée sous l'archontat de Glaucippe (3e année de la 92e Olympiade = 409 av. J.-C.); *Électre* est soit immédiatement antérieure, soit immédiatement postérieure à la tragédie homonyme d'Euripide, datée de 413. Enfin, l'opinion généralement admise situe les *Trachiniennes* entre *Antigone* et *Ajax*. En ce qui concerne *Philoctète*, on a pu émettre l'opinion que la date de composition était antérieure à la date de représentation. Voyez par exemple les introductions à chacune de ces tragédies dans l'édition de Dain et Mazon, « Collection des Universités de France », Paris (1955).

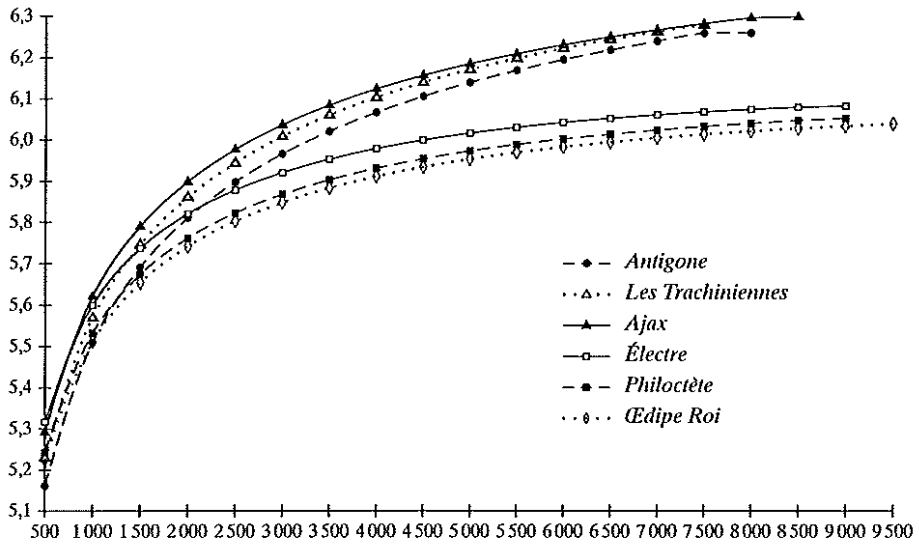

Fig. 1.– Entropie calculée

compris entre 1,12 et 1,19); la seconde regroupe *Œdipe Roi*, *Philoctète* et *Électre* (*B* est compris entre 1,21 et 1,25).

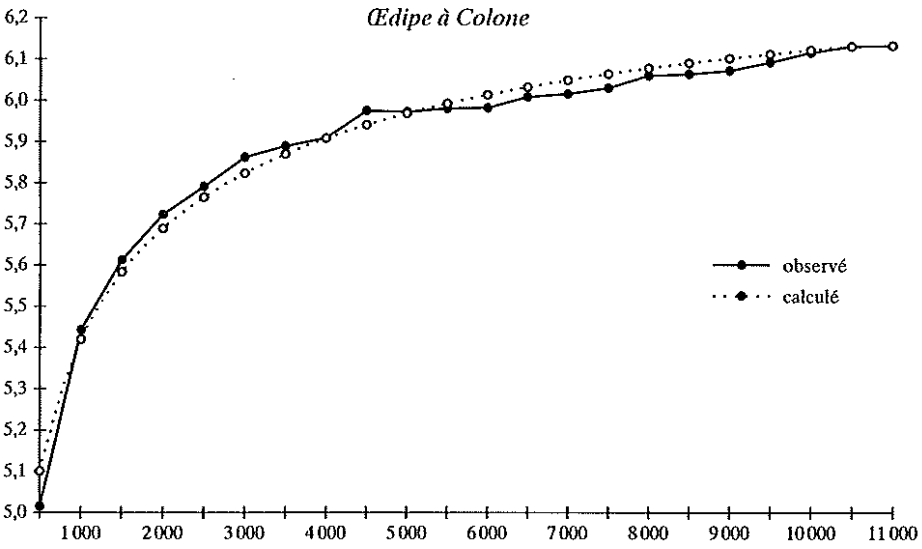

Fig. 2.– Coefficient *B* ; valeurs observées et calculées

La courbe dessinée par les valeurs trouvées pour *Œdipe à Colone* (figure 2) a le même aspect que celles dessinées par la première famille (correspondant à la valeur de $B = 1{,}14$), tandis que les valeurs successives (plus faibles) sont comparables à celles de la seconde.

On note aussi que la courbe des valeurs réellement observées pour les entropies est particulièrement régulière pour cette œuvre; tandis que pour les autres textes, il se dessine une sorte de « ressaut » aux environs de la moitié de la tragédie (*Antigone* : 4500–5000 [figure 3]; *Trachiniennes* : 3000–3500; *Ajax* : 3000–3500; *Œdipe Roi* : 2000–2500 ou 4000–4500; *Philoctète* : 4000–4500; *Électre* : 4000–4500).

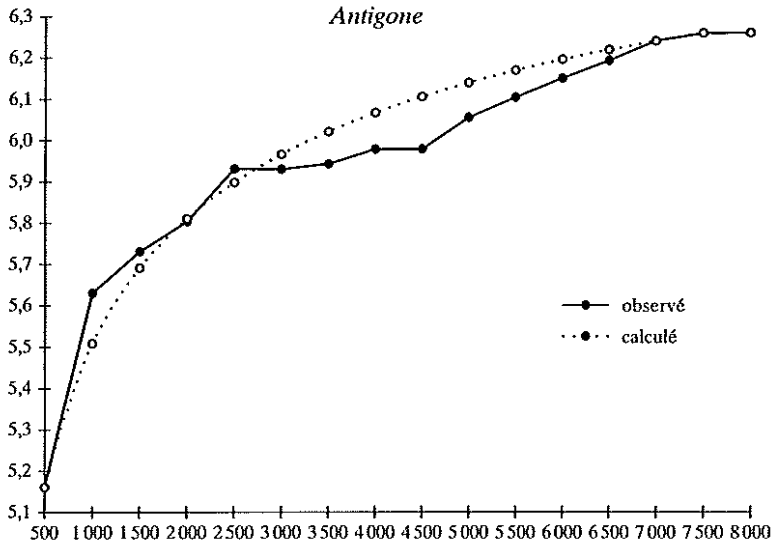

Fig. 3.– Coefficient B ; valeurs observées et calculées

Rôles des mots outils et de l'article

Il est possible, en considérant le code de catégorie, de décomposer les observations par catégories grammaticales. On observe alors que les courbes dessinées en ne retenant que les mots signifiants (substantifs, adjectifs, verbes, adverbes) et ceux des mots-outils qui sont des conjonctions de subordination ou de coordination sont presque parfaitement régulières. Il apparaît ainsi que les irrégularités de distribution sont, au moins pour une part, dues à l'article, aux particules (y compris les exclamations), et aux pronoms.

L'*article* (qui a toujours, on le sait, la fréquence d'emploi la plus élevée dans un texte quelconque) est d'un emploi quasi mécanique et obligatoire. Les *particules* et surtout les *exclamations* (ὦ πόποι, φεῦ, αἰαί) sont en nombre relativement restreint, et surtout d'emploi répétitif. L'influence des *pronoms*, de son côté, résulte sans doute du caractère narratif ou dialogué des textes.

De toute façon, ces deux dernières catégories impliquent des fréquences d'emploi élevées, et une forme particulière d'impossibilité de diversifier le vocabulaire. Toutefois, l'effet d'« essoufflement » provoqué par ces lemmes met sans doute en évidence — par contraste — une reprise de la diversification dans les autres catégories, au moment où la courbe se redresse.

Zones où la diversification est maximale

Le texte des tragédies a ceci de particulier qu'il est constitué de plusieurs unités ou même de genres littéraires différents (les dialogues, les tirades ou récits, les mélodrames, les chœurs) ; chaque genre pourrait avoir son vocabulaire particulier (voir par exemple l'emploi conventionnel du « dorien » dans les chœurs, mais aussi le vocabulaire volontairement « familier » de personnages comme celui du garde dans *Antigone*). On peut donc imaginer que le passage d'un genre à l'autre va se marquer par une diversification du vocabulaire.

Afin de mieux mettre en évidence les zones où l'enrichissement du vocabulaire est le plus manifeste, considérons les valeurs successives que prendraient les coefficients A et B si le texte se limitait à la portion examinée jusque là. On voit alors apparaître des alternances plus ou moins marquées de diminutions et d'augmentations de la valeur des coefficients A et B.

Ces variations se présentent très différemment d'une œuvre à l'autre; toutefois, elles ont toujours tendance à s'amenuiser quand on dépasse le premier tiers de l'œuvre.

Par conséquent, on peut considérer que la diversité du vocabulaire se marque surtout au début de l'œuvre (à ce moment, en effet, tous les mots sont d'abord considérés comme des *semel dicta*); l'apparition d'un nouveau mot au milieu et surtout à la fin influe moins nettement sur la valeur de A et de B.

Si on compare les fluctuations du coefficient B et le nombre des lemmes apparaissant pour la première fois dans chaque zone du texte (tableaux 2 et 3), en fonction de leur fréquence globale sur l'ensemble de l'œuvre, on observe deux séries de données manifestement liées. Les maxima du nombre de lemmes apparaissant dans chaque zone, et qui restent des *semel dicta* sur l'ensemble de l'œuvre, correspondent aux régions où la valeur du coefficient B diminue nettement par rapport aux valeurs précédentes.

Tableau 2

Valeurs successives de B et nombre de *semel dicta* par tranche

	Ajax		*Les Trachiniennes*		*Antigone*		*Électre*	
	B	*semel dicta*	*B*	*semel dicta*	*B*	*semel dicta*	*B*	*semel dicta*
500		83		69		46		62
1 000	1,174	69	1,114	91	0,968	89	1,143	109
1 500	1,029	107	1,150	70	1,132	62	1,211	66
2 000	1,071	73	1,189	60	1,182	65	1,301	37
2 500	1,112	66	1,240	46	1,145	98	1,326	42
3 000	1,154	57	1,292	38	1,195	50	1,311	53
3 500	1,152	77	1,249	97	1,220	59	1,324	53
4 000	1,132	93	1,221	93	1,220	65	1,321	41
4 500	1,138	60	1,209	77	1,239	45	1,255	100
5 000	1,136	83	1,174	108	1,202	95	1,244	57
5 500	1,139	72	1,156	87	1,183	86	1,252	39
6 000	1,150	70	1,143	93	1,164	94	1,251	45
6 500	1,165	52	1,139	80	1,148	108	1,247	54
7 000	1,161	91	1,145	61	1,128	104	1,253	37
7 500	1,168	73	1,156	56	1,126	69	1,256	39
8 000	1,189	39			1,126	4	1,267	20
8 500	1,189	18					1,254	64
9 000							1,254	31

Cette liaison apparemment paradoxale (si on suppose que le facteur B est un indicateur de diversification) s'explique, me semble-t-il, par l'évolution du calcul[10] de B.

En effet, quand un nouveau mot apparaît dans la zone examinée, la somme des entropies calculée sur les zones précédentes est augmentée d'une valeur très petite (la probabilité du nouveau lemme est égale à $1/N$, et le logarithme est par suite très faible). Simultanément, les probabilités de tous les mots employés antérieurement sont diminuées, puisque N augmente d'une unité. La somme des entropies les concernant diminue donc d'autant plus que la fréquence d'emploi d'un mot déjà rencontré est élevée. L'effet sur la somme totale est donc une variation négative.

[10] Dans les formules qui suivent, N représente le nombre total de formes examinées jusques et y compris la dernière zone considérée entièrement; n désigne le pas de progression, c'est-à-dire la valeur dont on augmente le nombre de formes prises en considération — si on examine par exemple un texte par pas successifs de 500 formes et si on rencontre un nouveau lemme à la 1 525^{e} forme, N vaut 1 500; n vaut 500.

Tableau 3

Valeurs successives de B et nombre de *semel dicta* par tranche

	Œdipe à Colone		*Œdipe Roi*		*Philoctète*	
	B	*semel dicta*	*B*	*semel dicta*	*B*	*semel dicta*
500		52		64		67
1 000	0,978	80	1,197	63	1,237	54
1 500	1,046	54	1,167	94	1,176	67
2 000	1,073	59	1,280	31	1,187	62
2 500	1,097	44	1,286	57	1,229	47
3 000	1,098	58	1,274	64	1,242	45
3 500	1,120	51	1,285	40	1,250	48
4 000	1,138	37	1,290	40	1,268	32
4 500	1,117	96	1,279	50	1,242	75
5 000	1,141	33	1,272	53	1,235	47
5 500	1,153	33	1,262	46	1,233	61
6 000	1,166	34	1,255	52	1,241	34
6 500	1,162	76	1,259	34	1,247	47
7 000	1,168	52	1,259	43	1,241	59
7 500	1,168	56	1,262	43	1,228	60
8 000	1,158	72	1,243	74	1,234	39
8 500	1,163	43	1,231	66	1,237	51
9 000	1,164	46	1,228	44	1,228	69
9 500	1,158	58	1,222	48		
10 000	1,149	64				
10 500	1,145	58				

En revanche, quand on examine la zone suivante du texte, si le même lemme réapparaît, les deux occurrences seront créditées d'une probabilité égale à $2/(N+n)$; le total sera donc augmenté d'une quantité égale à

$$\frac{4}{N+n} \times \left(2 \times \log \frac{2}{N+n}\right)$$

ce qui va accroître la valeur du total des entropies. De même, les lemmes qui avaient une fréquence élevée antérieurement, et qui sont à nouveau employés dans cette zone, vont voir leur fréquence augmentée, et celle-ci servira au calcul de l'entropie pour chacune de leurs occurrences. La valeur du total des entropies va donc augmenter, et par suite également le coefficient B.

Ainsi, quand la seconde zone considérée ne voit apparaître que peu de *semel dicta*, les augmentations du total des entropies correspondent à des valeurs de probabilité non seulement supérieures à $1/N$, mais supérieures aussi à ce qu'elles étaient pour ces mêmes mots dans les zones précédentes.

On voit que la multiplication des apparitions de *semel dicta* dans une zone déterminée correspond à une diminution de la valeur de B, par rapport à la valeur obtenue à la zone précédente (où un grand nombre de lemmes avaient déjà une fréquence supérieure à $1/N$), et est suivie d'une augmentation à la zone suivante (où un certain nombre des nouveaux lemmes obtiennent à leur tour une probabilité supérieure à $1/N$)[11].

On trouvera aux tableaux 4 et 5 les correspondances entre nombre de formes et références (n^{os} de vers) pour les zones ainsi définies, ainsi que les passages concernés pour chaque tragédie.

Tableau 4

	Formes	Vers	*B*	*S. dicta*
Antigone	2 000–2 500	341–439	1,182–1,145	98
	Premier chant du chœur — Premier récit du garde.			
	4 500–7 000	760–1 248	1,239–1,128	487*
	Le chœur célèbre l'amour — Thrène d'Antigone Malédiction de Tirésias — Évocation de Dionysos Récit de la mort d'Hémon — Lamento de Créon.			
Œdipe à Colone	4 000–4 500	663–761	1,138–1,117	96
	Éloge de Colone par le Chœur — Supplications de Créon.			
	6 000–6 500	985–1 076	1,166–1,162	76
	Œdipe se justifie devant Créon Thésée intervient pour libérer Antigone et Ismène.			
	7 500–8 000	1 246–1 329	1,168–1,158	72
	Récit de Polynice.			
Les Trachiniennes	3 000–4 000	493–678	1,292–1,221	135**
	Le chœur célèbre la puissance de l'amour Récit du viol de Déjanire — Mission de Lichas Le chœur célèbre le retour d'Héraclès.			
	5 000–6 500	846–1 129	1,174–1,139	288***
	Récit du suicide de Déjanire — Lamento d'Héraclès.			

* Soit 95 (de 4 500 à 5 000), 86 (de 5 000 à 5 500), 94 (de 5 500 à 6 000), 108 (de 6 000 à 6 500) et 10 (de 6 500 à 7 000).

** Soit 38 (de 3 000 à 3 500) et 97 (de 3 500 à 4 000).

*** Soit 108 (de 5 000 à 5 500), 87 (de 5 500 à 6 000), et 93 (de 6 000 à 6 500).

[11] Il est évident que, du point de vue de la fréquence globale, un lemme qui apparaît pour la première fois vers la fin de l'œuvre ne peut guère parvenir à une fréquence globale élevée; c'est moins évident quand on examine une tranche proche du début de l'œuvre. On aurait donc, en comparant la fréquence globale des lemmes apparaissant dans une tranche déterminée, un indice signalant les lemmes à portée très localisée.

Tableau 5

	Formes	Vers	*B*	*S. dicta*
Ajax	3 500–4 000	618–721	1,152–1,132	93
	Tecmesse — Récit de la folie d'Ajax Le Choryphée — Ajax ne se souvient plus.			
	6 500–7 000	1 148–1 245	1,165–1,161	91
	Ménélas et Teucros discutent du sort d'Ajax.			
Œdipe Roi	1 000–1 500	161–258	1,197–1,167	94
	Lamentations du Chœur — Proclamation d'Œdipe.			
	7 500–8 000	1 203–1 286	1,262–1,243	74
	Thrène du Chœur après la sortie d'Œdipe Récit de la mort de Jocaste et de la mutilation d'Œdipe.			
Philoctète	1 000–2 000	157–335	1,237–1,187	67
	Mélodrame entre le chœur et Néoptolème Apparition de Philoctète.			
	4 000–4 500	643–738	1,268–1,242	75
	Philoctète se prépare à partir en compagnie de Néoptolème Le chœur célèbre sa délivrance.			
	5 000–5 500	818–900	1,235–1,233	61
	Philoctète s'endort après la crise de douleurs Le chœur s'apitoie sur son sort Philoctète se réveille et se dirige vers le bateau.			
	6 500–7 500	1 042–1 239	1,247–1,228	60
	Ulysse contraint Philoctète à l'accompagner Les adieux de Philoctète à son refuge Mélodrame entre Philoctète et le chœur.			
Électre	4 000–4 500	690–777	1,321–1,255	109
	Récit (mensonger) de la mort d'Oreste par le Précepteur.			

Les zones de diversification maximale semblent bien correspondre aux « moments forts » de l'action dans chaque tragédie, que ceux-ci soient réellement représentés sur scène ou qu'ils fassent l'objet d'un récit, et non aux « styles » différents de l'œuvre.

Les lemmes nouveaux qui apparaissent dans ces zones, et dont la majorité n'est employée qu'une ou deux fois, se rattachent généralement au thème particulier traité dans le passage considéré. On observe notamment une série de lemmes qui, tout en étant employés plusieurs fois sur l'ensemble de l'œuvre, ne se rencontrent que dans les environs immédiats de la zone considérée; on pourrait les définir comme des « *semel dicta* relatifs ».

Ainsi en est-il, par exemple, des termes se rapportant à la course de chars dans le récit du Précepteur dans *Électre*. On y trouve notamment des noms de

cités ou de peuples (Ἀθῆναι, Αἰτωλία, Ἀργεῖος, Ἀχαιοί, Βοιῶτος...) d'où sont venus les participants à la course; les mots désignant les chevaux de course et leur équipement (ἅρμα, ἡνία, ὄχος, ἄξων, ἁρματηλάτης, διφρηλάτης, ζῦγον...); enfin, l'adjectif ἱππικός, employé sept fois sur l'ensemble, se rencontre six fois dans ce passage et le terme πῶλος, employé cinq fois sur l'ensemble du texte, ne se trouve qu'ici.

On constate le même phénomène dans l'éloge du dème natal de Sophocle au début de l'*Œdipe à Colone* : on y rencontre nombre de noms propres de divinités — Athéna, Aphrodite, Dionysos, Céphise; de noms géographiques — Ἀσία, Ἑλλάς; ainsi que plusieurs lemmes formés par composition avec le préfixe εὐ- : εὐήρετμος, εὐγένης, εὐθάλασσος, εὔιππος, εὔπωλος, εὐλογέω... ou encore des mots typiquement « élogieux » comme μυριόκαρπος, μητρόπολις, παιδότροφος qui ne se rencontrent pas dans le reste de l'œuvre.

On peut ainsi me semble-t-il considérer que les zones de diversification du vocabulaire marquent les épisodes spécialement importants de l'action, ou encore les compositions voulues par l'auteur comme des « morceaux de bravoure » — un peu à la manière des « grands airs » de l'opéra italien.

Notons aussi la structure particulière de l'*Antigone*, où la seconde moitié de l'œuvre — au vocabulaire particulièrement « riche » — voit s'intensifier, presque sans interruption, la tension dramatique que la première partie ne faisait que préparer.

Diversité du vocabulaire et composition de l'œuvre

En regroupant les tragédies d'après la forme générale de la courbe de variation de A, on voit se former deux groupes : l'un (figure 4), comprenant les *Trachiniennes*, *Antigone*, *Électre* et *Œdipe à Colone*, se compose d'une phase « modulée » en deux ou trois montées suivies de descentes marquant la diversité (cf. ci-dessus); à noter que, dans les *Trachiniennes*, la courbe redescend nettement à partir de la deuxième moitié de l'œuvre. Notons que la courbe des variations observée dans l'*Œdipe à Colone* est particulièrement régulière, avec une seule inflexion entre les formes 4 000 et 4 500.

Le second groupe (figure 5) présente dès le début (à partir des 1 000 premières occurrences) une chute suivie de remontées moins marquées. Ce groupe reprend *Ajax*, *Philoctète* et *Œdipe Roi*.

En comparant les courbes dessinées par les *Choéphores* d'Eschyle, et par l'*Électre* d'Euripide, on constate que la courbe de l'*Électre* de Sophocle présente le même aspect que celle d'Euripide, tandis que les *Choéphores* ont une distribution comparable à celles d'*Ajax*, de *Philoctète* et d'*Œdipe Roi*.

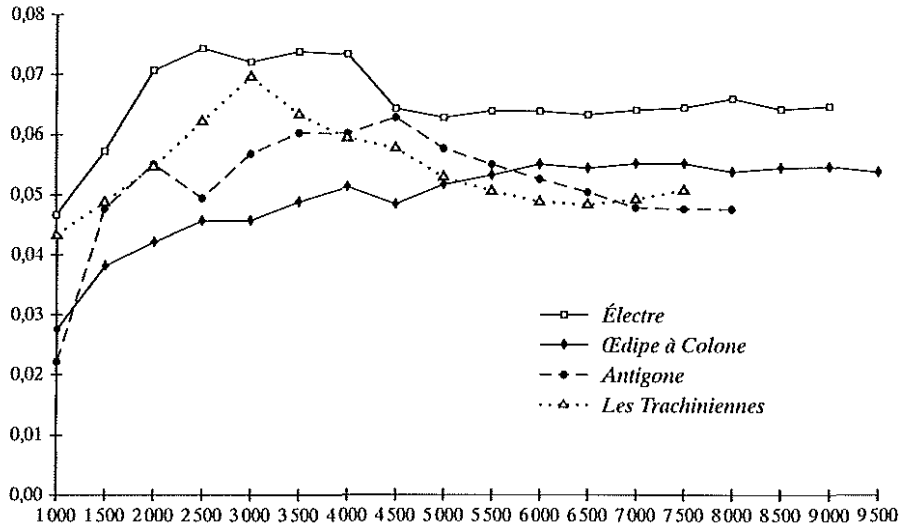

Fig. 4.– Variations du facteur *A*

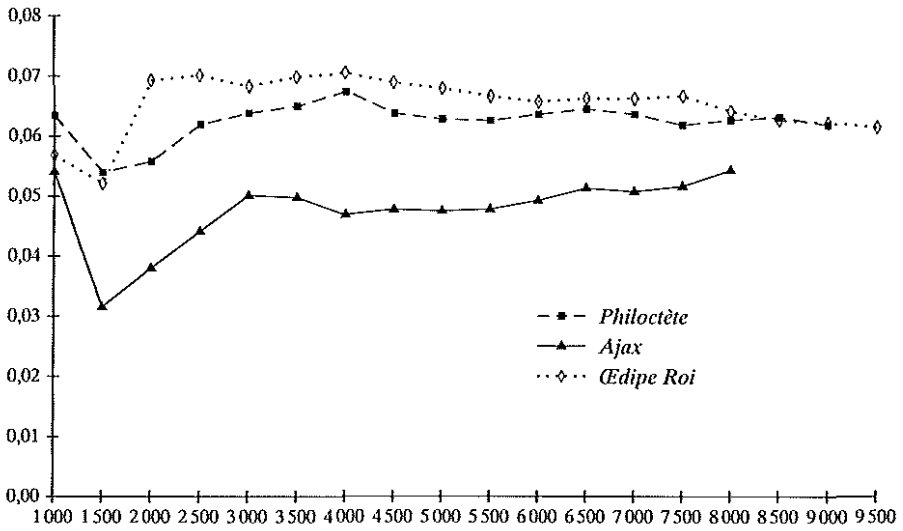

Fig. 5.– Variations du facteur *A*

Cette forme de distribution est apparemment typique d'une structure ancienne de la tragédie (qui, sous cette forme, tient plus de l'oratorio que de la

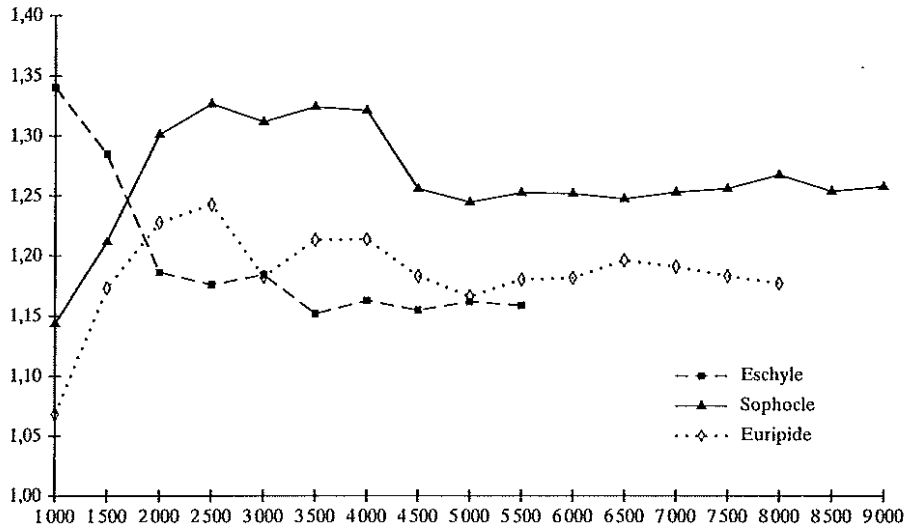

Fig. 6.– Les trois *Électre*; variations du facteur *B*

représentation « théâtrale »), où l'action se déroule sans rebondissements, une fois le « prologue » achevé.

Diversification au fil de la lecture

On peut se demander ce qui se passerait si on lisait le texte en ordre inverse (en commençant par le dernier mot). C'est ce qui a été fait au prix d'une très légère modification du programme.

Dans ces conditions, les valeurs extrêmes sont comprises, pour A, entre $-0{,}04$ (*Ajax*) et $-0{,}059$ (*Antigone*); pour B, entre 1,066 et 1,233 pour les mêmes œuvres. On observe que les courbes d'ajustement ainsi dessinées sont comme l'image réfléchie dans un miroir de celles obtenues en lisant dans l'ordre normal. Ceci se remarque très nettement en comparant les deux courbes obtenues à partir de l'*Antigone* (valeurs correspondant aux seuils de 2 000–2 500 et de 4 000–4 500).

On obtient pour A et B des valeurs différentes de celles observées dans l'ordre normal, ce qui paraît indiquer une relation des coefficients A et B avec le « moment » où le vocabulaire se diversifie le plus (c'est-à-dire au « début » ou à la « fin » du texte).

On observe également que les variations des valeurs observées vont en s'amenuisant au fur et à mesure de l'allongement du texte considéré, tout comme pour la lecture en ordre normal.

Le classement des œuvres par référence au coefficient B obtenu après lecture en ordre inverse paraît très différent du classement obtenu par la lecture en ordre direct. Toutefois, si l'on calcule la somme algébrique des coefficients A et B, et que l'on classe les œuvres en ordre croissant des valeurs obtenues en lecture directe, on voit réapparaître les deux groupes définis plus haut par les familles de courbes théoriques (*Antigone*, *Œdipe à Colone*, *Les Trachiniennes* et *Ajax* d'une part, *Œdipe Roi*, *Philoctète* et *Électre* d'autre part); dans chaque groupe, les valeurs obtenues par la lecture inverse vont en décroissant, tandis que les valeurs obtenues en lecture directe vont en croissant.

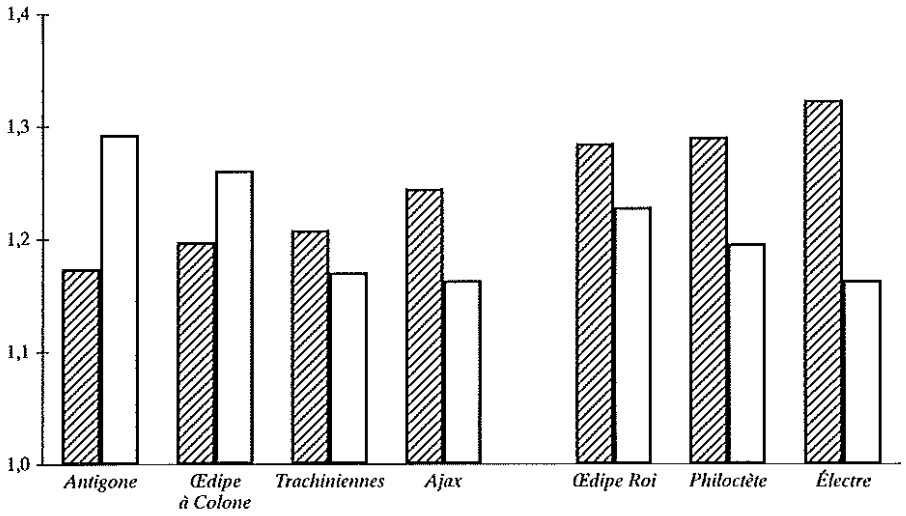

Fig. 7.– Somme des coefficients A et B ; lecture directe et inverse

De ce point de vue, *Œdipe à Colone* se rattache de manière évidente au premier groupe (valeurs intermédiaires entre *Antigone* et *Les Trachiniennes*). Il faut peut-être y voir deux périodes différentes de l'auteur, la première étant plus proche de la tragédie eschyléenne, comme semble l'indiquer l'évolution du coefficient B (voir ci-dessus), tandis que la seconde se rapproche des conceptions d'Euripide et de son « modernisme ».

Synthèse des observations

Les observations faites à partir du calcul de l'entropie d'un texte fournissent apparemment suffisamment d'indices, en ce qui concerne notamment la

chronologie des textes, ou les liens entre diversité du vocabulaire et composition de l'œuvre, pour justifier l'emploi de cette méthode dans une étude statistique d'une œuvre ou d'un groupe d'œuvres appartenant à une même époque ou à un même auteur.

Cependant, il va de soi que les indications obtenues de cette manière doivent toujours être rattachées à une observation directe des textes; elles servent plutôt à orienter l'analyse philologique qu'à fournir des explications fondamentales.

On s'attachera par exemple à relever les lemmes qui apparaissent pour la première fois dans les parties de texte indiquées par les divergences du coefficient B et le nombre de *semel dicta*; c'est là sans doute que se révèlent surtout les particularités du vocabulaire de l'auteur et de la structure du texte[12].

[12] On a vu que le nombre de lemmes nouveaux par tranche de 500 formes est plus élevé dans la tranche qui provoque une diminution temporaire du coefficient B. Or, presque tous les lemmes ainsi relevés ont une fréquence globale 1 ou 2, ou alors certains voient leur emploi limité à la zone où ils apparaissent.