Meta
Journal des traducteurs
Translators' Journal

META

Approche quantitative de l'étude des marqueurs du niveau de spécialisation dans les textes scientifiques anglais

Malcom Clay

Volume 34, Number 3, septembre 1989

1. Actes du Colloque Les terminologies spécialisées : Approches quantitative et logico-sémantique et 2. Actes du Colloque Terminologie et Industries de la langue

URI: https://id.erudit.org/iderudit/002342ar
DOI: https://doi.org/10.7202/002342ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Clay, M. (1989). Approche quantitative de l'étude des marqueurs du niveau de spécialisation dans les textes scientifiques anglais. *Meta*, *34*(3), 370–376. https://doi.org/10.7202/002342ar

APPROCHE QUANTITATIVE DE L'ÉTUDE DES MARQUEURS DU NIVEAU DE SPÉCIALISATION DANS LES TEXTES SCIENTIFIQUES ANGLAIS

MALCOM CLAY
Université Lyon 3, Lyon, France

Le problème de la reconnaissance des niveaux de spécialisation / vulgarisation dans des textes scientifiques peut être comparée à deux autres activités en sciences humaines utilisant couramment des approches quantitatives. La première est l'attribution d'un texte dont l'auteur est inconnu ou contesté, à l'œuvre d'un auteur donné pour qui on a établi les caractéristiques de ses œuvres incontestées. La deuxième est celle de la reconnaissance des formes — que ce soit celles des haches de l'époque de bronze, des caractéristiques des tombeaux dans un cimetière, ou la signification des différentes phases de la parade nuptiale des épinoches, pour ne citer quelques études assez connues dont la plupart ont utilisé comme moi l'*analyse factorielle des correspondances.*

Ma première approche a été quelque peu différente de celle que j'exposerai ici, en raison du corpus dont je disposais et de l'objectif que je m'étais fixé. Je disposais de textes écrits par le même auteur (professeur de physiologie orale) sur le même sujet (la fluorisation de l'eau) mais adressés à des publics différents. Mon intérêt premier se fixait sur la syntaxe de la vulgarisation scientifique en anglais, et mon objectif était de démontrer qu'il était possible de distinguer des textes de niveaux différents par leurs seules caractéristiques syntaxiques. Je suis donc parti de l'hypothèse contraire, postulant que les trois textes-témoins étaient syntaxiquement identiques et formaient un seul texte homogène, et j'ai soumis l'ensemble du corpus à l'analyse des correspondances.

Les premiers résultats des travaux ayant par ailleurs été publiés (Benzécri *et al., Pratique de l'analyse des données, 3. Langue, Linguistique, Lexicologie*, Paris, Dunod, 1981), je me contenterai de les commenter sous la forme de trois remarques liminaires, pour ensuite présenter quelques idées que j'ai développées par la suite, et qui me paraissent pouvoir alimenter la réflexion de tous ceux qui adoptent une approche quantitative des langues de spécialité.

Le corpus se compose donc de trois textes écrits par le même auteur sur le même sujet, mais destinés à des publics différents :

Texte A : chercheurs spécialistes
Texte B : dentistes praticiens
Texte C : grand public

Une première remarque terminologique s'impose. J'ai utilisé dans le titre le terme *niveau de spécialisation* pour référer à ces différents types de communication scientifique. Il ne me satisfait pas entièrement car il se focalise sur le texte, au détriment des variables situationnelles — comme si le texte avait une existence indépendante d'un émetteur, son auteur, et d'un destinataire, le *lecteur idéal* de Riffaterre, alors que les textes que j'étudie diffèrent essentiellement par le rapport entre leur destinataire et le

sujet traité. J'emploie l'expression *niveau de spécialisation* avec le souci d'éviter le terme *niveau de langue*, que je préfère réserver au sens que lui donne Bourquin, c'est-à-dire le rapport social entre locuteur et interlocuteur. Dans la typologie des actes de communication qui m'intéressent, le rapport entre émetteur et destinataire n'est pas *social*, mais induit par le rapport entre leur destinataire et le sujet traité. Le terme proposé par Mme Gallais-Hamono de *niveau de culture scientifique* conviendrait mieux, si on précisait *niveau de culture scientifique dans le domaine du sujet traité* pour éviter une possible confusion avec le niveau de culture scientifique en général.

Même en le précisant ainsi, nous n'arrivons pas à écarter l'inconvénient fondamental du terme *niveau* dont les collocations les plus fréquentes correspondent avec des mots de mesure comme *même, inférieur, supérieur*, portant le germe d'un jugement de valeur, et je l'utilise avec beaucoup de regret en espérant qu'un autre terme finira de s'imposer pour désigner les types de manifestations d'actes de communication qui se distinguent par les variables situationnelles concernées.

La deuxième remarque porte sur la nature du niveau *intermédiaire (Texte B)* entre les deux niveaux *extrêmes* de vulgarisation *(Texte C)* et de spécialisation *(Texte A)*. S'adressant à des dentistes praticiens, il sort un peu des classifications habituelles des niveaux de vulgarisation / spécialisation, qui ont tendance à inclure systématiquement, et souvent à privilégier, des manuels scolaires et universitaires. S'il est indéniable que des textes de cette nature occupent une place importante dans ce qu'on appelle les langues de spécialité, il importe de tenir compte des variables linguistiques introduites par la variable situationnelle qu'est le but — didactique en l'occurrence. La distinction entre *information scientifique* et *formation scientifique* est trop souvent passée sous silence dans les études de niveau en langues de spécialité.

La dernière remarque liminaire porte sur la façon d'envisager les rapports entre différents niveaux de spécialisation. Déformés par notre habitude de la linéarité des textes, nous avons spontanément tendance à penser qu'un texte *intermédiaire* va se situer quelque part entre les deux textes *extrêmes* — et ce *quelque part* va être sur une droite reliant les deux extrêmes.

Les figures 1 et 2 montrent clairement qu'il est nécessaire d'adopter une approche multi-dimensionnelle pour apprécier la distance qui sépare les textes : présenter cette distance sur une droite — même si elle comporte 78% de l'information résultant de l'analyse quantitative — revient à déformer la réalité de la distance séparant les textes, ce qui est particulièrement net pour la distance entre le texte de spécialisation *A* et le texte semi-spécialisé *B*.

Présenter les marqueurs du niveau de spécialisation / vulgarisation dans un colloque de terminologie en laissant de côté l'aspect lexico-sémantique peut prendre l'allure d'une provocation. Il n'en est rien. Cette hypothèse de travail avait sa justification propre — un exercice de style en quelque sorte, dont le but était de démontrer la nécessité de tenir compte de la dimension syntaxique dans les études de niveau.

Il peut être intéressant de se pencher sur les a priori qui président à cette démarche. Elle se justifie le plus souvent en affirmant que «tout le monde sait qu'il y a des différences lexico-syntaxiques entre texte de vulgarisation et texte de spécialisation». Dans une présentation pédagogique du problème, on citera des exemples très typés que tout le monde connait — et personne ne disputera l'appartenance de *œil au beurre noir* au niveau de vulgarisation et de *hématome bilatéral* au niveau de spécialisation. Mais souvent on franchit un pont de trop en laissant entendre que le terme *hématome périorbital* appartiendra exclusivement au(x) niveau(x) élevé(s), et que ce que son synonyme doit être rejeté pour ne figurer que dans le(s) niveau(x) bas. Or, tout traducteur sait pertinemment que ces synonymes rejetés peuvent très bien apparaître dans la communication écri-

te entre spécialistes, ne serait-ce que par ce qu'on peut appeler le *clin d'œil à un autre niveau* opéré dans le texte spécialisé, par exemple, par la citation de la lexie *(œil au beurre noir)* qui renvoie à la fois au concept et (surtout) au monde du quotidien.

On pourrait peut-être conclure que l'appartenance de *hématome périorbital* et de *œil au beurre noir* à leurs niveaux respectifs doit être formulée en termes de fréquence d'occurrence. Mais là encore, nous devons nous rendre compte que cette affirmation n'a qu'une valeur pédagogique, et que nous avons besoin de l'affiner pour qu'elle puisse avoir une valeur scientifique.

Pour illustrer mon point, je reviendrai à ma pratique de traducteur de langue de spécialité. J'avais demandé à mes étudiants de traduire en anglais un texte de vulgarisation scientifique dans le domaine de mes recherches, contenant le mot *carie* en français. Dans un premier temps, je m'apprêtais à vertement tancer ceux qui n'avaient pas perçu la différence entre la vision globale du mot anglais (*caries*, mais pas *à caries*) et la possibilité de vision globale ou de vision discrète en français (*une carie* ne relevant que de la vision discrète, donc ne correspondant pas à *caries*; *la carie* pouvant être, soit globale, donc correspondant à *caries*, soit discrète, donc ne correspondant pas à *caries*. Dans un deuxième temps, je m'apprêtais à démontrer que si on pouvait à la rigueur considérer que traduire *carie* (vision globale) par *caries* allait de soi dans un texte de spécialisation, il n'en était pas de même pour un texte de vulgarisation. Je tenais là un bon marqueur de niveau de spécialisation / vulgarisation, car l'intuition de l'anglophone écarte caries du niveau de vulgarisation, le réservant aux niveaux plus spécialisés. Je me suis donc tourné vers mon corpus pour illustrer mon propos. Mais à ma surprise, pour les trois textes correspondant aux trois niveaux, je relève des différences peu importantes dans le nombre d'occurrences du mot *caries*. Je ne peux pas affiner mon discours *pédagogique* en ajoutant le concept de fréquence d'occurrence — les trois textes sont de longueur égale (Table i.). Cette constatation heurte mon intuition linguistique d'anglophone, et je m'oblige à regarder les textes de plus près (Table 2).

Utiliser la fréquence d'occurrence d'un mot (ici le mot *caries*) comme critère descriptif (Table 1) est une démarche qui a le grave inconvénient, inhérent à toute approche probabiliste, de poser le problème en termes d'équiprobabilité d'occurrence. Or, la signification du nombre d'occurrences du mot *caries*, n'est pas à établir en fonction du nombre total de substantifs (et encore moins en fonction du nombre de mots, toutes catégories grammaticales confondues comme on le fait trop souvent). La probabilité de son occurrence n'a de sens que dans le contexte de l'ensemble des références constatées pour le même concept. Dans notre corpus, cette probabilité ne peut être estimée qu'en tenant compte en même temps des occurrences de *dental decay* et de *tooth decay* (Table 2). La véritable significationi du nombre d'occurrences n'apparaît que si l'on établit des catégories conceptuelles s'éloignant de l'immédiatement observable.

Un autre exemple tiré de l'utilisation d'un descripteur syntaxique servira à bien montrer cette nécessité de dépasser l'observable.

On lit souvent que l'utilisation de la voix passive est une des marques du discours de spécialiste. Cependant, la probabilité de trouver un emploi significativement plus fréquent des formes passives dans le discours du spécialiste ne se vérifie pas bien dans les faits, alors que ceci correspond à notre intuition linguistique et à nos attentes stylistiques. Si l'on compte le nombre de formes passives et on les rapporte au nombre de propositions dans chaque texte, on obtient un critère qui ne discrimine pas très bien le texte de vulgarisation du texte de spécialisation. La raison en est que là encore, on approche le problème en termes d'équiprobabilité. Mais la probabilité réelle de trouver exprimée à la forme passive la phrase *La terre tourne autour du soleil* est voisine de zéro, alors que la phrase Nous avons chauffé le corps à une température de 130°C est bien davantage sus-

ceptible d'apparaître à la forme passive. Il faut donc comparer des choses comparables et exprimer le nombre de formes passives par rapport au nombre de formes effectivement susceptibles d'être passivées, c'est-à-dire ne pas baser une approche quantitative sur l'immédiatement observable, les occurrences, mais sur les options réellement ouvertes et les choix faits par l'encodeur.

Ce deuxième cas illustre bien la nécessité pour le linguiste de construire des catégories conceptuelles pour l'étude quantitative, car dans le deuxième exemple ci-dessus, l'encodeur aurait pu effacer la référence au sujet *nous* non pas par l'utilisation de la forme passive, mais par l'utilisation de *on*. Notre catégorie immédiatement observable (utilisation des formes passives) doit s'élargir pour tenir compte de toutes les marques de la présence ou de l'absence de l'énonciateur dans son énoncé, dont l'utilisation de la forme passive n'est qu'une des manifestations.

Ces réflexions ne sont pas faites pour invalider toute tentative de dépouillement et d'analyse *automatique* des textes. C'est une approche nécessaire quand on est confronté à des documents de taille importante, et l'analyse *primaire* des données peut fournir des indications précieuses qui ne seraient pas accessibles ou facilement perceptibles par d'autres moyens. Cependant, pour l'analyse *fine* des données que requiert la discrimination entre niveaux de spécialisation, il me paraît indispensable d'éviter de traiter les occurrences en termes d'équiprobabilité sous-jacente, et souhaitable de créer des catégories conceptuelles tenant compte des choix faits par l'énonciateur.

Ce même raisonnement est susceptible de s'appliquer non seulement aux données à soumettre à une analyse quantitative, mais aussi à l'interprétation des résultats de l'analyse quantitative. Nous avons spontanément tendance à interpréter les résultats, soit en termes des unités ayant servi au dépouillement (les mots ou, dans le cas qui m'intéresse, les phrases), soit en termes des unités d'appartenance (les textes, les chapitres, ou dans le cas de ma recherche, l'article scientifique). Il en résulte, soit une présentation très atomisée des résultats respectant toute l'information recueillie, mais permettant mal de visualiser les différences significatives (Fig. 3), soit une vue très générale des différences entre les textes, très *parlant*, mais avec une perte d'information considérable (Fig. 2).

On pourrait penser que la solution consisterait à regrouper les données en unités intermédiaires entre la phrase et le texte — le paragraphe, par exemple. Cette solution, qui ne pose pas de problèmes techniques, se heurte cependant à deux objections théoriques. La première concerne la nature de l'unité intermédiaire. Les paragraphes, souvent de longueur inégale, ont l'inconvénient d'être des unités qui n'ont pas toujours été choisies par l'auteur et qui peuvent avoir été imposées après la remise du manuscrit, pour des raisons purement matérielles. Cette première objection n'est pas suffisante pour un rejet définitif du paragraphe comme unité intermédiaire servant à l'interprétation et à la présentation des résultats. Cependant, une deuxième objection théorique plus importante doit nous inciter à plus de prudence. Interpréter et visualiser des différences en utilisant comme unités des phrases, des paragraphes, ou des textes, présente l'inconvénient d'assimiler le processus d'appréhension à un processus statique que nous saisissons à des points fixes. Il me semble plus raisonnable de postuler que la perception par le lecteur de l'appartenance du texte qu'il est en train de lire à un niveau de spécialisation / vulgarisation donné est un processus dynamique qui évolue avec chaque nouvel apport. Je propose donc qu'il est souhaitable d'aborder le problème en faisant appel à d'autres unités. J'en présenterai deux.

La première approche *dynamique* pourrait consister à visualiser par le point *«t»* les informations recueillies au moment *«t»* de la lecture. Ainsi :

◆ le point A1 représenterait la somme des informations recueillies après la lecture de la première phrase du texte A,

◆ le point A2, la moyenne de la somme des informations contenues dans les deux premières phrases,

◆ le point A3, la moyenne des informations contenues dans les trois premières phrases... et

◆ le point A65 représenterait pour un texte de 65 phrases, la moyenne des informations contenues dans le texte entier, le «centre de gravité» du texte, identique au point «A» de la figure 1.

D'autres approches *dynamiques* pourraient:

◆ effacer le poids des informations anciennes en visualisant des séquences (de 10 phrases, par exemple) dont on représenterait le centre de gravité par un calcul de moyennes mobiles (Fig. 4)

A10 = (A01 + A02 + A03 + A04 + A05 + A06 + A07 + A08 + A09 + A10) / 10
A11 = (A02 + A03 + ... + A10 + A11) / 10
A12 = (A03 + A04 + ... + A11 + A12) / 10
...
A65 = (A56 + A57 + ...+ A64 + A65) / 10

◆ privilégier davantage l'apport des informations nouvelles provenant de la phrase qui vient d'être lue en pondérant celle-ci par son numéro d'ordre:

A01 = A1 / 1
A02 = (A1 + A2*2) / 3
A03 = (A1 + A2*2 + A3*3) / 6
...
A65 = (A1 + A2*2 + ... + A65*65) 2145

En conclusion, tant pour l'analyse des données que pour l'interprétation de l'analyse, deux approches du texte, utilisant deux types d'unités de compte, sont souhaitables. La première approche est basée sur un balayage de l'observable — affixation, formes passives, longueur et place des circonstants, etc., par exemple, et une interprétation de l'analyse basée sur des unités de compte observables — les mots, les phrases, par exemple. Cette première démarche a l'avantage de préjuger le moins possible de ce que l'on va trouver, et peut ouvrir des perspectives insoupçonnées de nouvelles lectures du texte.

La deuxième approche, qui fait suite à la première et lui est complémentaire, évite de traiter les occurrences en termes d'équiprobabilité sous-jacente. Basée sur des catégories conceptuelles créées par le linguiste, elle tente de cerner les choix faits par l'énonciateur. Pour l'interprétation des résultats, elle adopte une démarche heuristique, et ne se limite pas aux unités de compte immédiatement observables, mais tente d'envisager les phénomènes interprétés comme relevant d'une démarche dynamique de la part de l'encodeur, comme du décodeur.

Notre espoir est que ce type d'approche quantitative permettra de dépasser la description exhaustive des marqueurs pour aborder le problème de l'évaluation de leur poids relatif dans le repérage du niveau de spécialisation.

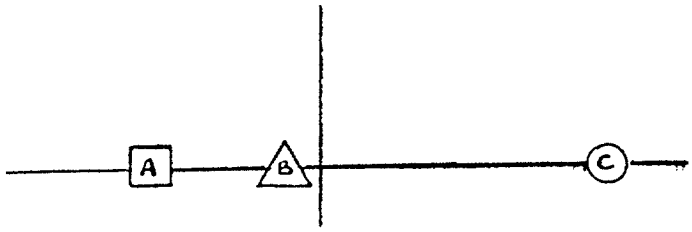

Fig. 1 Distances entre les centres de gravité des textes (distances AB, AC, et BC) sur axe 1 qui comporte 78% de l'information, apportée par la longueur et/ou la complexité des constituants immédiats de la proposition.

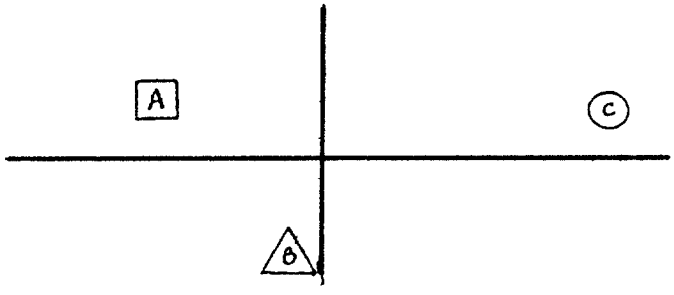

Fig. 2 Distances entre les centres de gravité des textes (distance AB, AC, et BC) sur axes 1 et 2. L'axe comporte 19% de l'information, apportée par la complexité de la structure propositionnelle de la phrase.

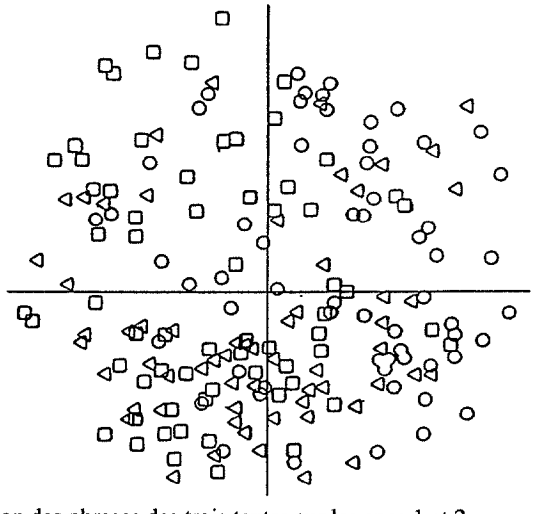

Fig. 3 Représentation des phrases des trois textes sur les axes 1 et 2.

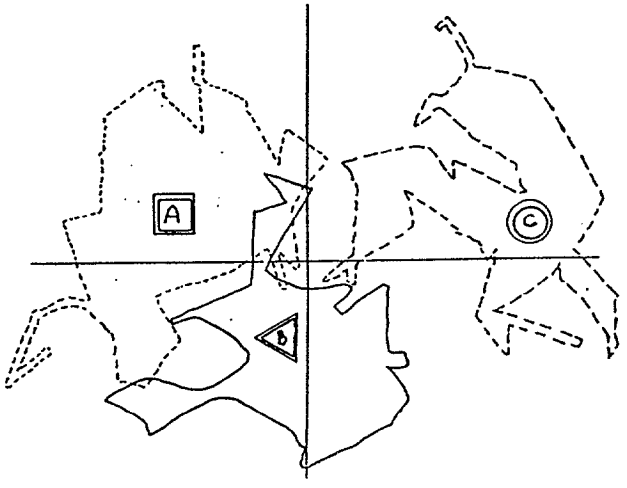

Fig. 4 Aires des textes établis à partir des points correspondant aux séquences de dix phrases. Cette approche *dynamique* semble indiquer qu'un petit nombre de séquences du texte C (vulgarisation) sont perçues comme ayant des caractéristiques (ici syntaxiques) voisines d'un petit nombre de séquences du texte B, et vice-versa. Par contre, certaines séquences du texte B pourraient être perçues comme très voisines de certaines séquences du texte A (spécialisé), et d'autres séquences de ces deux textes pourraient être perçues comme ayant les mêmes caractéristiques.

Table 1. Occurrence de *caries*

	A	*B*	*C*	*TOTAL*
CARIES	17	14	12	43
AUTRES	2710	2829	2977	8516
TOTAL	2727	2843	2989	8559

Table 2. Occurrence de caries / dental decay / tooth decay

	A	*B*	*C*	*TOTAL*
CARIES	17	14	12	43
DENTAL DECAY	6	24	14	44
TOOTH DECAY	6	16	27	49
TOTAL	29	54	53	136