Document generated on 07/18/2025 2:19 p.m.

Meta
Journal des traducteurs
Translators' Journal

META

Mesure des relations lexico-sémantiques dans des textes scientifiques : problèmes méthodologiques

Nathan Ménard

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/003629ar
DOI: https://doi.org/10.7202/003629ar

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

Ménard, N. (1989). Mesure des relations lexico-sémantiques dans des textes scientifiques : problèmes méthodologiques. *Meta*, *34*(3), 468–478. https://doi.org/10.7202/003629ar

Tous droits réservés © Les Presses de l'Université de Montréal, 1989

This document is protected by copyright law. Use of the services of Érudit (including reproduction) is subject to its terms and conditions, which can be viewed online.

https://apropos.erudit.org/en/users/policy-on-use/

érudit

This article is disseminated and preserved by Érudit.

Érudit is a non-profit inter-university consortium of the Université de Montréal, Université Laval, and the Université du Québec à Montréal. Its mission is to promote and disseminate research.

https://www.erudit.org/en/

MESURE DES RELATIONS LEXICO-SÉMANTIQUES DANS DES TEXTES SCIENTIFIQUES : PROBLÈMES MÉTHODOLOGIQUES

NATHAN MÉNARD
Université de Montréal, Montréal, Canada

0. DE LA LEXICOMÉTRIE À LA SÉMANTIQUE QUANTITATIVE

L'analyse de la structure et de la typologie textuelles a toujours accordé une grande importance au vocabulaire, et non sans raison : les unités lexicales sont le lieu privilégié où se manifestent les choix des auteurs tout en reflétant par ailleurs des contraintes de toute nature. Elles constituent le point d'appui des études de caractéristiques, des tentatives d'attribution (hypothèses d'individuation du style), mais donnent lieu aussi à des calculs de constantes pour des textes d'un même ensemble, indépendamment des auteurs (hypothèses fonctionnalistes et typologistes).

Les études quantitatives fondées sur les mots définis de manière purement formelle — avec ou sans lemmatisation — ont abouti à des comparaisons éclairantes à cet égard pour la stylistique, la sociolinguistique et la psycholinguistique (on connaît bien les travaux de l'école mullérienne et leurs applications). Des publications récentes portant sur le traitement des réponses libres dans des enquêtes socio-économiques et autres (Lebart et Salem, 1988) montrent bien que la lexicométrie, fortement soutenue par des moyens informatiques à sa mesure, peut encore exploiter des veines très riches. Force est de constater, toutefois, que la prise en considération des problèmes sémantiques dans ces calculs a été souvent restreinte sinon totalement évacuée. La raison principale ? Ce n'est certainement pas par manque d'intérêt des données sémantiques : les nombreuses analyses de contenu,de sémiotique et de linguistique textuelle prouveraient le contraire. Il faut invoquer plutôt le fait que toute tentative de mesure se heurte à la subjectivité inévitable des analyses sémantiques. C'est donc un problème de méthode, et, comme souvent avec la sémantique, il est plus commode de l'ignorer ou de le contourner plutôt que d'y faire face, selon l'expression de Kerbrat-Orecchioni (1977) : «La sémantique ou comment s'en débarrasser».

Quand on en vient aux descriptions statistiques, les problèmes sont encore plus complexes : comment prétendre faire le décompte d'entités aussi peu tangibles que les sèmes, les noèmes, les traits ou les primitifs sémantiques ? L'objet à calculer ne semble a priori réductible ni à des unités discrètes ni à un continuum segmentable. Néanmoins, à défaut d'aborder directement ces unités, on peut tenter de rendre plus explicites certaines hypothèses ou plus modestement certaines intuitions sur la structure sémantique des textes : la notion de cohésion textuelle est à notre avis mesurable par le biais d'un calcul des relations lexico-sémantiques. Un tel calcul peut également contribuer à mieux caractériser des sous-ensembles de textes — en ce qui nous concerne ici, des textes de vulgarisation scientifique.

Nous attirerons l'attention sur les problèmes méthodologiques posés par 1) le corpus, 2) les classes lexico-sémantiques, 3) les indices statistiques.

1. DES TEXTES SCIENTIFIQUES AUX OUVRAGES DE VULGARISATION

Dire d'un texte donné (livre, article, note) qu'il s'agit d'un texte scientifique suppose qu'on y a reconnu ou qu'on lui attribue a priori certaines caractéristiques dont la présence ou l'absence le distingue des autres types de texte. Mentionnons quelques-uns :

a) La fonction didactique — d'où une sélection fine de l'audience, souvent désignée explicitement dans la préface.

b) La condition d'objectivité — il s'agit de présupposés d'énonciation largement imposés par le positivisme : l'auteur d'un discours scientifique répond de la vérifiabilité ou de la «falsification» de ses énoncés soit par nécessité théorique et déduction logique, soit par expérimentation ou observation, soit par référence à une autorité ou à une source reconnue. Zones grises : les sources théocratiques,la Genèse contre Darwin.

c) La nouveauté ou l'originalité relative du propos, en ce sens que l'information, de quelque nature qu'elle soit (données, analyse, méthode d'investigation, critique), répondrait à des questions ou à des problèmes qui se posent aux récepteurs.

En ce qui concerne les propriétés linguistiques — donc une fois considéré le texte clos, dégagé des conditions d'énonciation — on peut encore signaler :

d) Un emploi plus strict de la terminologie — décelable par l'usage restreint des synonymes ou d'équivalents dans la référence aux «objets» de connaissance et, par le fait même, une plus grande licence par rapport aux contraintes rédactionnelles qui concernent la réitération ou la répétition.

e) Une répartition assez nette du vocabulaire en trois sous-paradigmes :

- Le vocabulaire métalinguistique, qui peut-être aisément repéré grâce à sa concordance à la liste du *Vocabulaire général d'orientation scientifique*. On y retrouve des termes généraux d'exposition et d'argumentation, des connecteurs stylistiquement marqués, des termes de référence intratextuelle ou cotextuelle (ADMETTRE, SUPPOSER, HYPOTHÈSE, SOIT, SI...ALORS, FIGURE CI-CONTRE, TABLEAU SUIVANT, ETC.).
- Les sous-ensembles terminologiques, repérables soit par des définitions d'auteur (par RELATIVITÉ RESTREINTE nous entendons...etc) soit par leur inclusion dans des lexiques spécialisés.
- Le sous-ensemble de termes communs qui s'explique non seulement par l'usage normal, aléatoire, du vocabulaire fondamental mais aussi, en particulier dans les textes de vulgarisation scientifique et les manuels, par le souci d'exemplification et d'illustration.

On peut encore faire appel à d'autres indices pour reconnaître les textes scientifiques, comme le taux de monosémie ou d'oligosémie — qui est forcément en corrélation avec un vocabulaire hautement spécialisé — ou comme l'importance relative du co-texte en tenant compte non seulement des graphiques, tableaux, schémas et formules mais aussi des phrases qui y font référence, parfois de façon indépendante de la syntaxe du texte. En effet même si un roman est abondamment illustré, il n'est pas dans les conventions d'y trouver des phrases du genre «Le héros assomma le méchant comme on le voit dans le dessin ci-contre».

En apportant les corrections nécessaires aux conditions c) et d) ci-dessus, on peut dire que ce qui vaut pour la reconnaissance des textes scientifiques en général s'applique aussi à des sous-ensembles de ceux-ci, dont les textes de vulgarisation. Et parmi ces derniers, ceux destinés à un public universitaire — mais non spécialiste — et qu'on retrouve par exemple dans les collections *Point Science, Science ouverte* ou *Petite bibliothèque Payot*, occupent sans doute une place distincte dans la typologie. À en juger par le nombre de titres récents dans toutes les disciplines, il s'agit bien d'un sous-ensemble de

plus en plus important et qui répond au besoin de donner accès à des champs de connaissances cloisonnés par la structure des programmes universitaires. Ils sont différents des manuels en ce sens que la fonction didactique est augmentée d'un dessein *autodidactique*, d'où peut-être un recours plus accentué à des formes d'explication analogique ou comparative.

Sans prétendre en faire ici la démonstration, nous considérons que les conditions ou les propriétés que nous venons d'exposer permettent de définir les textes de vulgarisation pour universitaires, et que font partie de cette catégorie des ouvrages comme *les Grandes Économies* de Barou et Keizer, *le Calcul, l'imprévu* de Ekeland, le *Macrocospe* de Rosnay, la *Relativité* d'Einstein (traduit par Solovine), *Patience dans l'azur* de Reeves. Nous appliquerons des calculs de relations lexico-sémantiques à des extraits d'Einstein et de Reeves, avec l'*Exil et le Royaume* de Camus comme point de comparaison.

2. LES RELATIONS LEXICO-SÉMANTIQUES

Une grille opérationnelle des relations lexico-sémantiques est un outil indispensable à l'étude de la cohésion des textes. Les travaux de Halliday et Hasan, de Gutwinsky, de Patry et Ménard ont montré l'intérêt de telles analyses, même si les trois premiers auteurs ont attaché beaucoup plus d'importance à la cohésion dite grammaticale. Nous partons du principe que la cohésion est crée par toute une série de liens sémantiques qu'un locuteur parvient à établir entre des paires de mots d'un bout à l'autre du texte.

Ces liens auraient pu être assimilés entièrement aux fonctions lexicales de Mel'cuk. Mais dans la pratique, notamment dans l'analyse des textes scientifiques, on se trouve devant des associations de diverse nature et pour en rendre compte, la définition de ces fonctions demanderait plus de souplesse, d'autant plus que les informations sur les choses (de nature encyclopédique et souvent non prévisibles par le lexicographe) constituent en général l'essentiel du contenu de ces textes. Il n'en demeure pas moins que c'est la sémantique structurale qui nous a fourni les quatre grandes classes de relations que nous utiliserons, avec leurs subdivisions. (voir tableau à la fin).

2.1 LES RELATIONS D'ÉQUIVALENCE

Elles concernent les paires de synonymes et de coréférents lexicaux. En général on tient compte aussi des anaphores et cataphores (pronoms, possessifs, certains déictiques), ce qui permet entre autres de corriger la fréquence relative des éléments thématiques, et donc de faire une évaluation plus réaliste des mots-thèmes et des mots-clés — c'est la subdivision 1.1.1 du tableau, nous pouvons l'ignorer pour le moment.

Les problèmes soulevés par les classes d'équivalence ne manquent pas. La distinction nécessaire entre synonymie de langue et synonymie de discours conduit ici à des impasses (Cruse 1988, Patry et Ménard 1988), et seule une voie moyenne — non exempte d'arbitraire — permet de prendre des décisions satisfaisantes. Ainsi MASSE MANQUANTE et MATIÈRE NOIRE sont des synonymes dans Reeves simplement parce que, dans le texte, ils sont coréférentiels. La décision de les classer en 1.1.2 se trouve renforcée par le fait que l'auteur les utilise systématiquement comme des syntagmes figés ou en voie de l'être, ou encore par leur valeur sémantique dans un champ notionnel structuré (ici une terminologie) qui sert de cadre d'interprétation à ce discours. Si BÊTE et MOUTON sont en relation d'hyperonymie (1.2.1) c'est aussi grâce au principe de référence actuelle, même si dans leur cas les définitions courantes des dictionnaires ne laissent pas beaucoup de choix. Mais entre ÉTOILE et SOLEIL, ÉTOILE et NAINES BRUNES, seule une définition contextuelle permet de trancher en faveur de l'équivalence.

Les relations entre certains dérivés avec suffixes modificateurs et leurs bases peuvent aussi relever de la substitution asymétrique, s'il n'y a pas changement de classème.

La nature et l'importance de la modification peuvent toutefois obscurcir la relation. CAMIONNETTE demeure un (PETIT) CAMION, et d'ailleurs au Québec le deuxième terme est plus souvent employé, parfois sans l'adjectif, pour désigner ce type de véhicule. LIVRET et LIVRE seront classés également en 1.2.2 même si la modification est plus complexe : LIVRET ne désigne que certains types de petits livres. Mais en cas de doute, c'est la condition de coréfentialité, même considérée avec beaucoup de souplesse, qui permet d'établir les équivalences, les cas de référents distincts étant récupérables par les autres classes.

Sur ce problème se greffe un autre non moins important, celui de la délimitation et du regroupement des mêmes unités de vocabulaire — cas de réitération. La norme Muller appellerait ici quelques adaptations. Si l'on est sensible au statut syntagmatique de MATIÈRE NOIRE, VOIE LACTÉE, en se fondant sur les nomenclatures spécialisées, on est un peu perplexe devant NAINES BRUNES ou NÉBULEUSE D'ANDROMÈDE à cause des micro-séries ÉTOILES NAINES, NAINES ROUGES, ou GALAXIE D'ANDROMÈDE, SATELLITES D'ANDROMÈDE, NÉBULEUSE DU CRABE, NÉBULEUSE D'ORION, qui nous incitent plutôt à segmenter de telles expressions, ce qui reviendrait à appliquer la norme Muller sauf dans les cas vraiment figés.

Quant à savoir si deux occurrences d'un même lexème forment une même unité de vocabulaire, la règle idéale serait de ne réunir que les cas où il y a identité de référence, étant donné les conséquences sur la reconnaissance d'autres classes de relations. Prenons les exemples suivants de Reeves (p.29)

- «Le soleil est une étoile, semblable aux milliers d'étoiles que nous apercevons la nuit à l'œil nu.»
- «Fermez les yeux. Après quelques minutes, ouvrez-les sur la voûte étoilée.»

La première occurrence de ÉTOILE a une valeur distincte de la deuxième qui a le sens courant de «tout astre visible, excepté le soleil et la lune» (Petit Robert) ; et l'adjectif ÉTOILÉE dérive de ÉTOILE-2 et n'a rien à voir avec le soleil. Pour être strict, il y aurait des relations hyponymiques entre SOLEIL et ÉTOILE-1 ou entre ÉTOILE-2 et ÉTOILE-1, des relations de même niveau taxinomique (3.2.1) entre SOLEIL et ÉTOILE-2. Dans cet exemple, la distinction des référents coïncide avec des différences de sens dans les dictionnaires.

Ce type de «collision» polysémique n'est pas rare dans les textes de vulgarisation scientifique. Il faudrait normalement en tenir compte. Et si, comme c'est le cas dans la présente étude, on décide de l'ignorer pour des raisons pratiques (coût du traitement, beaucoup de cas limite) il convient quand même de s'assurer, à partir d'un petit échantillon, que la fréquence de ces phénomènes peut être négligée sans grand risque de fausser l'interprétation des résultats.

2.2 LES RELATIONS MORPHO-LEXICALES

Nous avons déjà évoqué le cas des dérivés qui n'entraînent pas un déplacement de la classe sémantique principale du mot de base — disons que les deux ont encore le même «genre prochain». Leur traitement n'est pas compliqué, pas plus que celui des affixés qui changent de catégorie, à condition que l'on puisse retrouver sans problème le sens du mot de base. Quelquefois l'évolution peut masquer quelque peu l'étymon commun : GALAXIE/LAIT/VOIE LACTÉE. Mais l'idée même de texte scientifique nous autorise à établir ces liens même lorsque les étymons sont considérés comme distincts : EAU/HYDRAULIQUE, LOIN/TÉLESCOPE. Ainsi, en thermodynamique par exemple, le lien sémantique entre ENTROPIE et TRANSFORMATION (de chaleur en énergie mécanique) est difficile à établir ; on peut toutefois s'appuyer sur les liens étymologiques

que les auteurs établissent dans les textes mêmes : «...du grec *entropê* signifiant changement» (de Rosnay).

Néanmoins il faut un minimum de «motivation» pour que l'existence d'une relation soit reconnue, même si la dérivation formelle peut être clairement établie. De même on peut s'interroger sur les liens entre le terme simple ANNÉE et ses composés ANNÉE-LUMIÈRE, ANNÉE GALACTIQUE. Dans le premier cas il y a bien préservation du sens de ANNÉE, même s'il y a changement de classème du composé («distance» au lieu de «durée»). Mais, pour être une unité de temps, ANNÉE GALACTIQUE n'en est pas moins une formation analogique qui n'a rien de comparable à notre période (même approximative) de douze mois. Beaucoup de lexies complexes posent le même problème par rapport à leurs composants. Aussi longtemps qu'on parviendra, par chaîne analogique ou par implication, à justifier la motivation sémantique, on maintiendra le lien. Sinon, il faut y voir un facteur de perturbation dans les calculs de cohésion, le récepteur-lecteur pouvant être amené à créer de fausses relations.

2.3 LES CONTRASTES

Entre la définition traditionnelle des antonymes (réanalysés en 3.1 en quatre sous-classes) et les valeurs distinctes à l'intérieur d'une même série taxinomique, il y a des nuances, ou plutôt des degrés dans les oppositions. Ce que toutes les paires de mots retenues en 3. ont en commun, quelle que soit la sous-classe à laquelle elles appartiennent, c'est la présence d'au moins un sème contraire ou contradictoire et dont la pertinence (ou la valeur informative) dans le texte ne saurait être négligée. À côté des oppositions déjà codées dans le lexique (NAÎTRE/MOURIR, NOCTURNE/DIURNE) il y a celles établies ou précisées par le texte (À PIED/À CHEVAL, GÉNÉRAL/RESTREINT). Des problèmes se posent avec les sous-classes *tout/partie* (3.1.5) et les oppositions par synecdoque. De même, les oppositions multiples (3.2) dont la force associative est diluée, contrairement aux binaires, risquent d'échapper à l'analyste. Si la terminologie scientifique est assez bien structurée, la consultation d'un lexique permet de prévoir ou de récupérer plusieurs de ces relations.

Enfin si nous avons réuni les oppositions cycliques et orthogonales dans une même sous-classe, ce n'est pas par confusion. Leur rendement combiné autant en langue qu'en discours nous paraît relativement faible et justifie ce regroupement.

2.4 LES COLLOCATIONS ET IMPLICATIONS DIVERSES

Ces relations peuvent être considérées comme plus lâches que les autres. D'ailleurs c'est une classe qui récupère la plupart des associations qui n'ont pu être clairement ramenées à une subdivision quelconque des classes précédentes. Cela pose déjà le problème d'un traitement plus raffiné des résidus. On y trouve toutefois un dénominateur commun : ce sont des rapports syntagmatiques qui unissent les paires de mots retenus. Dès qu'il y a une contrainte distributionnelle qui a pour effet que tel prédicat appelle automatiquement tel argument, que tel régissant est associé à tel subordonné, le lien sémantique se juxtapose au rapport syntaxique. Un CHEVAL peut bien trébucher ou faire un faux pas, mais c'est le terme BRONCHER qui est comme réservé à l'expression de ce fait. De même, à l'intérieur d'un texte de physique, la probabilité de trouver THÉORIE quand on parle de RELATIVITÉ est très forte.

Il n'est pas difficile non plus de reconnaître les relations de présupposition lexicale (implication par définition). Quant à ce que nous appelons implications indirectes ou médiatisée, ce sont des relations dont nous avons conscience mais qui sont sans doute trop liées aux faits d'expérience, de culture ou de situation, bref trop contingentes pour que la sémantique générale les intègre sans risque. Or il se trouve que dans les textes, ces

types d'associations sont légion. Il peut s'agir de relations métonymiques: ÉCLAIR/-FOUDRE, NUIT/LUNE. Einstein structure ses exemples avec la série TRAIN, VOYAGEUR, TALUS, WAGON, SERRER, FREINS, SECOUSSE, etc., des mots qui entrent tous dans un champ d'expérience banal et une organisation matérielle sans surprise, et qui par conséquent aident à la cohésion du texte. Par ailleurs, quand c'est l'auteur lui-même qui établit une relation d'implication par définition entre UNIFORME et CONSTANT, entre TRANSLATION et MOUVEMENT, nous n'avons guère le choix. Parfois il faut trouver un troisième terme qui fasse la jonction: entre PRÉTENDRE et VÉRITÉ il y a FAUSSETÉ qui est en relation contrastive avec le dernier et est un présupposé du premier.

Mais il y a toujours des cas limite. Entre CAILLOU et PARLER (texte de Camus) fera-t-on le lien en souvenir de certaines pratiques d'art oratoire?

3. STATISTIQUES ET INDICES

Appliquons maintenant cette grille à l'analyse au chapitre 5 d'Einstein(«le principe de la relativité au sens restreint»), et au chapitre 7 de Reeves («l'architecture de l'Univers — le monde des étoiles»). Les données pour Camus sont tirées d'une étude de quatre tranches de 500 mots de «l'Hôte» (Ménard 1987), et dont nous tirons des moyennes.

3.1 LES VALEURS

Soit un texte $T[a+g+a+b+f+...+z]$, dans lequel *a, b,* etc sont les mots retenus pour l'analyse, et R_i la série de relations lexico-sémantiques que ces derniers entretiennent entre eux, nous aurons les valeurs suivantes (dont quelques-unes sont courantes en statistique lexicale):

- N: le nombre de mots du texte (occurrences).
- V: le nombre de vocables (au sens mullérien) du texte, c'est-à-dire le total de mots différents.
- $\mathbb{V}$: le nombre de vocables considérés comme mots référentiels en excluant les mots dits grammaticaux (articles, préposition, etc.). Remarquez toutefois que dans le calcul des anaphores et de la cohésion grammaticale, les pronoms et la plupart des mots dits grammaticaux sont comptés.
- $\$$: le nombre de vocables ayant chacun au moins une relation lexico-sémantique quelconque *(R)* avec un autre vocable du texte. $\$$ est un sous-ensemble de $\mathbb{V}$.
- R_L: le nombre de relations lexicales (réalisées) entre les éléments de $\$$ pris deux à deux.
- $F_{(X)}$: la fréquence d'un élément ou d'un ensemble donné.

Pour les fins de la présente étude, un même vocable n'est compté qu'une fois dans $\$$, même si on relève chacune des relations qu'il peut entretenir avec les autres. Ainsi FORCE n'entrera qu'une seule fois dans $\$$, mais dans R_L il y aura toutes ses relations (avec ÉNERGIE, ACCÉLÉRATION, MASSE, INERTIE, ETC.). Le nombre total de relations varie entre zéro (cas limite ou de non-texte) et un maximum (texte très redondant, la saturation est atteinte lorsque tous les vocables sont reliés les uns aux autres sans exception). Le nombre maximum de relations peut être aisément calculé par algèbre combinatoire

$$R_{max} = \frac{\$!}{2!(\$ -2)!}$$

Il est possible à partir de ces valeurs de calculer plusieurs indices simples (des rapports) dont

$\frac{\$}{V}$: qui mesure la densité ou la spécificité de cohésion en rapport avec la richesse lexicale, en ce sens que pour deux textes comparables en fonction de V ou de $\mathbb{V}$, cet indice fera ressortir lequel exploite davantage la cohésion lexico-sémantique.

$\frac{F(\$)}{V}$: indice qui mesure de façon plus précise la densité de la cohésion, en tenant compte de la répétition des éléments de $. Dans certains travaux (Halliday-Hasan, Patry) la répétition pure et simple constitue une sous-classe de relations cohésives, et on ne saurait en nier l'importance. Il nous a semblé préférable, pour faciliter les calculs statistiques, de récupérer les effets de la répétition grâce à cet indice, ce qui revient au même en fin de compte. Mais de toute manière nous ne l'utiliserons pas pour le moment.

Si deux textes affichent des valeurs égales pour l'un ou l'autre de ces indices simples, on considérera comme le plus cohésif celui dont les vocables ont en moyenne le taux le plus élevé de relations mutuelles. Ce qui nous conduit à l'indice composé de cohésion :

$$\mathcal{C}_L = \frac{\$}{V} \times \frac{log\, R_L}{log\, R_{max}}$$

l'échelle logarithmique sert à corriger l'effet de la progression de R_{max}.

3.2 INTERPRÉTATION DES RÉSULTATS (voir tableau statistique)

Même si la plus grande prudence s'impose au stade actuel de notre expérience — on doit se rappeler qu'il s'agit de résultats obtenus sur des extraits — on peut au moins faire quelques observations.

a) À longueur égale ou comparable (N≈500), nos textes de vulgarisation scientifique sont nettement plus faibles en richesse lexicale (V) que le récit, Reeves dépasse largement Einstein-Solovine, même si ce dernier texte a un accroissement plus fort de N=500 à N≈1000. Reeves évite davantage la répétition et recourt dans ses explications à une gamme assez large d'exemples et d'images résolument anthropomorphiques qui font augmenter le sous-ensemble de termes communs.

b) La distribution des R_i est assez intéressante en soi. D'abord en valeur absolue, le total des R_L chez les scientifiques est au moins trois fois supérieur à celui de Camus. Les valeurs relatives de chaque classe (%) font voir aussi des points de démarcation entre les deux types de texte : chez les premiers, recours massif aux dérivés et composés pour soutenir la progression thématique. En revanche pour les équivalences, Reeves se rapproche de Camus, ce qui renforce l'observation précédente à propos de la répétition.

Les données relatives des classes 3 et 4 ne sont pas concluantes, ou du moins ne signalent pas de différences marquées. On aurait pu s'attendre à un pourcentage plus élevé de relations contrastives chez les scientifiques, mais les extraits considérés semblent relever d'un discours argumentatif-descriptif plutôt que de la réfutation. Néanmoins ils conservent leur supériorité en valeur absolue.

Enfin on peut émettre certaines réserves quant aux effectifs de la classe 4. Les «implications indirectes» finissent par être le point de chute de bon nombre de résidus et de relations de toutes sortes. Une analyse plus fine s'impose pour qu'on puisse tirer des conclusions sérieuses.

c) Quant à la cohésion proprement dite, la valeur des indices (simple ou composé) est environ deux fois plus élevée dans les textes de vulgarisation scientifique que dans le récit et l'extrait d'Einstein-Solovine domine les deux autres. On ne spéculera pas sur ces écarts ; ces chiffres renforcent simplement nos impressions à la lecture de ces textes.

Par ailleurs la stabilité relative de ces indices dans chaque texte est un très bon signe, même si l'on doit au moins constater leur tendance à augmenter avec la longueur du texte.

d) Reste le problème de l'évaluation des écarts en probabilité. En ce qui concerne la distribution des R_L, le test de Kolmogorov-Smirnov nous semblerait tout indiqué ; il permettrait de tenir compte non seulement des écarts entre deux textes pour chaque classe mais aussi du cumul des différences. Or le classement des relations n'étant pas tout à fait étanche, le calcul serait donc plus conforme à la nature des faits. Seul problème ici : il vaudrait mieux avoir plus de 4 classes pour obtenir de meilleures conditions d'application du test.

Les indices eux-mêmes doivent encore être soumis à de nombreuses vérifications et expériences avant qu'on puisse établir un seuil de signification des écarts. C'est une tâche difficile et fastidieuse mais pas impossible.

CONCLUSION

Des trois séries de problèmes méthodologiques dont nous venons de discuter, l'analyse et le classement des relations lexico-sémantiques demeurent encore les plus complexes. Toutefois, en prenant des risques calculés, on parviendra quand même à établir des statistiques qui jettent un éclairage un peu plus objectif sur quelques-unes des propriétés d'un texte. Avec les réserves que nous avons émises et qui appellent autant que possible des corrections pour la poursuite des expériences, les indices de cohésion reflètent assez bien la spécificité des textes de vulgarisation scientifique.

TABLEAU DES RELATIONS LEXICO-SÉMANTIQUES

1. **ÉQUIVALENCE** (même catégorie syntaxique)
   - 1.1 **Permutables**
     - 1.1.1 anaphores et cataphores(pronoms,possessifs, déictiques).
     - 1.1.2 synonymes et coréférents lexicaux
       Ex. : MATIÈRE NOIRE / MASSE MANQUANTE
   - 1.2 **Substituables** (asymétriques)
     - 1.2.1 hyperonyme / hyponyme
       Ex. : BÊTE / MOUTON
       ÉTOILE / SOLEIL, NAINES BRUNES
     - 1.2.2 base / base + modificateur
       Ex. : SALUT / SALUTATION
       CAMION / CAMIONNETTE
       MILLE / MILLIER, MILLIARD

2. **RELATIONS MORPHO-LEXICALES** (avec changement de classème)
   - 2.1 **Dérivés** (avec ou sans changement de catégorie syntaxique)
     - 2.1.1 de même base
       Ex. : VOIR / VISIBLEMENT
       LUMIÈRE / LUMINEUX,LUMINOSITÉ
       LAIT / VOIE LACTÉE

2.1.2 de bases différentes (mais même sémène)
Ex.: EAU/HYDRAULIQUE
LACTÉE/GALAXIE
LOIN/TÉLESCOPE

2.2 **Composés**
2.2.1 élément/lexie composée
Ex.: DYNAMIQUE/ÉLECTRO-DYNAMIQUE
ANNÉE/ANNÉE-LUMIÈRE
2.2.2 élément/lexie complexe
Ex.: GALAXIE/ANNÉE GALACTIQUE

3. **CONTRASTE**
3.1 **Binaires**
3.1.1 graduables
Ex.: LENTEMENT/VIVEMENT
GÉNÉRAL/RESTREINT
3.1.2 non graduables
3.1.2.1 en opposition privative
Ex.: NAÎTRE/MOURIR
MOUVEMENT/INERTIE
3.1.2.2 en opposition équipollente
Ex.: À PIED/À CHEVAL
NOCTURNE/DIURNE
THÉORIQUE/RÉEL
3.1.3 converses
Ex.: QUESTION/RÉPONSE
3.1.4 en opposition directionnelle et antipodale
Ex.: EN ARRIÈRE/EN FACE
3.1.5 tout/partie
Ex.: ATOME/ÉLECTRON, NOYAU
3.2 $X_{\text{-naire}}$ (x>2)
3.2.1 de même niveau taxinomique
Ex.: DIVAN/CHAISE
COLLINE/PLATEAU
NÉBULEUSE D'ANDROMÈDE/VOIE LACTÉE
3.2.2 sériés
Ex.: MÉTHANE/ÉTHANE/PROPANE
3.2.3 en opposition cyclique et orthogonale
Ex.: EN ARRIÈRE/À COTÉ
PARALLÈLE/PERPENDICULAIRE

4. **COLLOCATIONS ET IMPLICATIONS DIVERSES**
4.1 **Prédicat/argument, régissant/subordonné**
4.1.1 collocation contraignante
Ex.: GERMAIN/COUSIN
NUCLÉIQUE/ACIDE
4.1.2 collocation à forte probabilité
Ex.: BRONCHER/CHEVAL
CHAMP/(DE) GRAVITATION
THÉORIE/(DE LA) RELATIVITÉ

4.2 **Implication par définition**
(certaines relations de présupposition lexicale)
Ex.: RECOMMENCER / ARRÊT
NASEAU / CHEVAL
ESTUAIRE / FLEUVE

4.3 **Implication indirecte, «médiatisée»**
Ex.: SAVOIR / JOURNAL
CAILLOU / PARLER(?)
ÉCLAIR / FOUDRE

TABLEAU STATISTIQUE ET INDICES / Les % sont entre parenthèses

	EINSTEIN (et Solovine)			REEVES			CAMUS
	1e tranche / 2e tranche		Total	1e tranche / 2e tranche		Total	Moyenne
N	500	439	939	500	87	87	500
V	173	169	283	212	219	324	238,8
	137	130	249	177	183	278	187,0

R_L et %

1. Equi.	12 (0,06)	15 (0,09)	26 (0,06)	28 (0,15)	35 (0,17)	58 (0,17)	8,8 (0,15)
2. Morph	66 (0,36)	47 0,29)	142 (0,35)	51 (0,28)	47 (0,23)	108 (0,32)	6,8 (0,11)
3. Contr	37 (0,20)	32 (0,20)	87 (0,21)	33 (0,18)	45 (0,22)	51 (0,15)	14,5 (0,24)
4. Coll. Impl.	71 (0,38)	67 (0,42)	151 (0,37)	71 (0,39)	78 (0,38)	121 (0,36)	29,3 (0,49)
TOTAL RL	186	161	406	183	205	338	59,4
$	103	105	214	122	121	197	79,5

INDICES

$ / V	0,752	0,808	0,859	0,689	0,661	0,709	min 0,37 max 0,417
C_L	0,4568	0,4783	0,5985	0,4030	0,3958	0,4148	min 0,2008 max 0,2119

RÉFÉRENCES

CRUSE, D.A. : *Lexical Semantics*, Cambridge University Press, 1987
GUTWINSKY, W : *Cohesion in Literary Texts*, Janua Lingua, Mouton, The Hague, 1976.
HALLIDAY, M.A.K and HASAN, R. : *Cohesion in English*, Longman, 1976
KERBRAT-ORECCHIONI, C. : *De la sémantique lexicale à la sémantique de l'énonciation*, thèse, Université de Lille III, 3 tomes, 1977
LEBART, L. et SALEM, A. : *Analyse statistique des données textuelles*, Dunod, 1988
MARTIN, R. : *Inférence, antonymie et paraphrase*, Klincksieck, 1976
MEL'CUK, I. et coll : *Dictionnaire explicatif et combinatoire du français contemporain*, PUM, 1984
MÉNARD, N. : *Mesure de la richesse lexicale*, Slatkine-Champion, 1983
MÉNARD, N. : «Calcul de la cohésion lexico-sémantique des textes : aspects méthodologiques et recherche d'indices statistiques !, *Actes du colloque de l'ALLC*, Oxford University Press, 1987
MULLER, C. : *Principes et méthodes de statistique lexicale*, Hachette, 1977
PATRY, R. : *Le lexique dans l'analyse de la cohésion linguistique*, thèse, U de M, 1986
PATRY, R. et MENARD, N. : «La synonymie de la langue est-elle celle du discours» polycop., U de M, 1988

TEXTES

CAMUS, A. : *L'exil et le royaume*, Gallimard, Folio, 1980
EINSTEIN, A. : *La relativité*, traduit par M. Solovine, Payot, 1986
REEVES, H. : *Patience dans l'azur*, Seuil, 1988
ROSNAY (de) J. : *Le macroscope*, Seuil, 1975