Meta
Journal des traducteurs
Translators' Journal

META

Pour un index évolutif et cumulatif de cooccurrents en langue techno-scientifique sectorielle

Philippe Thoiron et Henri Béjoint

Volume 34, numéro 4, décembre 1989

URI : https://id.erudit.org/iderudit/004488ar
DOI : https://doi.org/10.7202/004488ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Thoiron, P. & Béjoint, H. (1989). Pour un index évolutif et cumulatif de cooccurrents en langue techno-scientifique sectorielle. *Meta*, *34*(4), 661–671. https://doi.org/10.7202/004488ar

POUR UN INDEX ÉVOLUTIF ET CUMULATIF DE COOCCURRENTS EN LANGUE TECHNO-SCIENTIFIQUE SECTORIELLE

PHILIPPE THOIRON ET HENRI BÉJOINT
Université Lumière Lyon 2, Lyon, France

INTRODUCTION

L'existence d'affinités entre les mots est une réalité évoquée depuis longtemps par de nombreux linguistes (Sweet 1899, Bally 1951, Firth 1957, etc.) Pour Sauvageot (1964 : 69), une énorme partie du vocabulaire est constituée de mots à rayon d'action limitée : «Ces vocables ne se construisent qu'avec un nombre plus ou moins réduit d'autres mots, toujours les mêmes, avec lesquels ils forment des locutions stéréotypées». Dans un texte plus récent, Lafon *et al.* (1985 : 68) soulignent eux aussi l'importance de ce phénomène :

> Ce qui est au bout de la langue ou qui vient au bout de la plume, ce n'est pas le mot tout seul, c'est une certaine quantité de mémoire. Un mot n'arrive jamais nu. Appendus à lui et invisibles, tous les autres collent à lui à divers degrés. Nous n'avons dans la tête que des agglutinations (...) nées du moirage de ces habitudes de locutions héritées.

Ces affinités engendrent un phénomène de cooccurrence pour certains mots qui forment ainsi, sur la chaîne syntagmatique du discours, des groupements préférentiels.

La connaissance de ces groupements est indispensable à la maîtrise de la langue, surtout à partir d'un certain degré de compétence (Benson 1985 : 189). En particulier, c'est elle qui permet au traducteur de bien choisir entre des synonymes dont il doit connaître les restrictions d'emploi en fonction des contextes[1]. On peut citer, à la suite de Sinclair (1966 : 429), le cas de la série anglaise *vigorous, energetic et forceful* dans laquelle seul *vigorous depression* forme un ensemble qui fonctionne dans le domaine de la météorologie. Ceci ne pose aucun problème au décodage, mais en pose par contre à l'encodage (Mackin 1978 : 150 ; Cowie 1981 : 225). Le traducteur peut aussi être démuni devant les variations possibles en fonction de l'appartenance à la catégorie grammaticale. C'est ainsi que l'anglais aura côte à côte *desperate need* et *desperately need* mais que *badly need* ne correspond pas à **bad need* (Greenbaum 1974 : 81 ; voir aussi Mackin 1978 : 150-153).

Mais il demeure difficile, sinon impossible, de trouver ces groupements dans la lexicographie ou la terminographie actuelle, même pour des langues aussi étudiées que le français ou l'anglais. Pour l'anglais, Benson (1985) cite les ouvrages de Reum (*Dictionary of English Style*, 1955), de Rodale (*Word Finder*, 1947), de Katsumata (*Kenkyusha's New Dictionary of English Collocations*, 1958) et de Friederich et Canavan (*Dictionary of English Words in Context*, 1979) avant de mentionner son propre *BBI Combinatory Dictionary : A Guide to Word Combinations in English* (Benson *et al.* 1985). On peut y ajouter Dzierzanowska et Douglas Kozłowska (1982). Pour le français, on peut citer Lacroix (1937) et Gak (1963). Le but de cet article est d'examiner certains

des problèmes posés par la rédaction d'un recueil de ces groupements, conçu comme outil d'encodage destiné, en particulier, au rédacteur et au traducteur spécialisés.

On a choisi, dans le cadre du travail exploratoire présenté ici, d'éviter la langue littéraire dont une des caractéristiques est précisément de s'affranchir des groupements trop prévisibles. C'est un des moyens, pour l'auteur, d'atteindre à l'originalité de l'expression. La langue techno-scientifique, au contraire, évite l'originalité (cf. à cet égard les instructions normatives draconiennes parfois données aux auteurs par certaines revues scientifiques ou techniques) et recherche souvent une expression aussi «neutre» que possible. Bref, ce n'est pas le lieu privilégié des effets de style.

NOTION GÉNÉRALE DE COOCCURRENCE

La notion générale de cooccurrence repose sur le fait que, d'une manière générale, les mots ont des affinités particulières, ainsi qu'en témoignent les phénomènes de prévisibilité sur la chaîne syntagmatique. Ces affinités sont essentiellement de deux types :

- affinités seulement sémantiques (cf. «échos» au sens de Lafon 1984) ;
- affinités syntactico-sémantiques : les mots entretiennent des relations sémantiques étroites à l'intérieur d'une structure syntaxique. Ceci a été vu par de nombreux linguistes, notamment Sweet (1899) et Hockett (1958), même si pour certains (Halliday 1966, Sinclair 1966, Crystal et Davy 1969) l'aspect strictement lexical l'emporte. À cet égard, un titre comme celui de l'article de Meier (1975) : «On Placing English Idioms in Lexis and Grammar» est révélateur.

Seul le deuxième type d'affinités peut faire l'objet d'un recueil utile au traducteur, soucieux d'évaluer le degré de contrainte ou de liberté, au plan syntaxique comme au plan sémantique, lié à l'emploi d'un terme donné.

Ces affinités syntactico-sémantiques se manifestent à l'intérieur d'ensembles qui peuvent aller du mot composé à la phrase en passant par le syntagme, lexicalisé ou non. Mots composés et syntagmes lexicalisés peuvent être organisés le long d'un continuum, mais ils relèvent de la lexie ; ils sont déjà consignés dans les dictionnaires et les répertoires terminologiques — même si cela pose des problèmes pratiques lorsqu'il s'agit de leur trouver un article d'accueil (cf. Meier 1975 : 223) — et ne sont donc pas concernés par la présente étude. Sont exclus également les idiotismes (définis ici comme groupements préférentiels de mots, dont le sens global n'est pas équivalent à la somme des sens de leurs éléments ; définition d'ailleurs traditionnelle — cf. Falk 1973 : 36-37, Hockett 1958 : 172, Robins 1964 : 70), quel que soit leur niveau d'intégration dans la hiérarchie syntaxique de la phrase ; on négligera aussi bien l'idiotisme de phrase (ex. : «Quand le chat n'est pas là, les souris dansent») que l'idiotisme intraphrastique (ex. en anglais : *to kick the bucket* ou ex. en français : «tirer son chapeau»). Il s'agit toujours de groupements dont le caractère figé — pétrifié au sens de Leech 1981 — interdit toute manipulation : on a bien *kick the bucket*, mais **kick the pail* est impossible (Cowie 1983 : xii).

CRITÈRES DE SÉLECTION DES GROUPEMENTS PRÉFÉRENTIELS

Il y a deux types de critères de sélection de ces groupements en vue de la rédaction d'un recueil :

- ◆ les critères syntactico-sémantiques
- ◆ les critères quantitatifs

I. CRITÈRES SYNTACTICO-SÉMANTIQUES

De ce point de vue, on peut considérer que les frontières entre constituants syntaxiques sont d'autant plus infranchissables que les constituants sont plus haut dans une hiérarchie dont la phrase est le sommet. En termes d'analyse en constituants immédiats,

les constituants de phrase sont alors considérés comme autonomes. La phrase est vue comme potentiellement formée de trois éléments :

- le(s) complément(s) de phrase ;
- le sujet ;
- le prédicat.

Tous ces constituants ne sont pas nécessairement présents en même temps. Si, pour des raisons de commodité d'analyse, l'accent était mis délibérément sur l'homogénéité syntaxique des groupements, on pourrait exclure d'emblée tous ceux qui contiendraient des éléments n'appartenant pas tous au même constituant de phrase. Il s'agirait d'une décision susceptible de conduire à des conséquences ponctuelles fâcheuses, notamment dans le cas de l'anaphore ou de l'ellipse : tel élément mentionné à un endroit de la chaîne dans un constituant donné peut être absent dans un constituant ultérieur (ex. : «Il adorait les timbres. Il faisait d'ailleurs une collection depuis l'âge de sept ans.» *Collection de timbres* est un groupement privilégié dont les éléments ne sont pas présents, dans ce cas-là, au sein d'un même constituant de phrase.) Greenbaum insiste bien d'ailleurs sur cet aspect éventuellement transphrastique : «*Collocates need not be in the same sentence and can even cross utterances by different speakers*» (1974 : 80). Ceci peut être pallié de deux façons :

- il existe, ailleurs dans le texte, d'autres occurrences de ce groupement ;
- si ce n'est pas le cas, le groupement devra être reconstitué à partir d'informations prises hors du corpus.

Dans cette approche, on favorise délibérément la surface des énoncés. Il peut même y avoir franchissement des frontières de constituants de la structure profonde de la phrase, si la structure résultante est tout entière intégrée dans un même constituant de phrase en surface. C'est ainsi que des nominalisations comme *le ronflement du moteur* sont prises en compte en dépit de la relation sujet-prédicat sous-jacente. Les groupements «nom + adjectif épithète» sont eux aussi enregistrés, même si on peut, en structure profonde, voir entre les éléments des relations de type prédicatif (Lees 1960 ; Vendler 1968).

Une délimitation radicale des groupements en fonction de critères strictement syntaxiques présente des avantages si on envisage un traitement automatique du texte par un analyseur syntaxique capable d'isoler très rapidement les grands constituants de la phrase. Cependant, les palliatifs évoqués ci-dessus ne fonctionnent pas en toutes circonstances. C'est ainsi, par exemple, que la nominalisation n'est pas possible avec certains emplois verbaux (ex. : «L'eau gèle en arrivant au sol» -> «*le gel de l'eau...») ou que, si elle est possible, elle pose des problèmes de choix lexico-sémantiques (ex. : «fondre» -> «la fonte»/«la fusion»). Il semble donc que, pour la délimitation des groupements, l'imposition de frontières placées entre les trois grands constituants syntaxiques d'une phrase soit plutôt néfaste. Pour des raisons pratiques et au stade expérimental où se situe un travail comme celui-ci, on peut toujours être amené à respecter la frontière de phrase (sans méconnaître, évidemment, les inconvénients de cette décision).

Peut-on, dès lors, envisager la phrase dans son ensemble comme unité lexicographique d'un recueil ? Au plan plus strictement sémantique, dans un modèle de production linguistique (Sens-Texte de I. Mel'čuk par exemple), on peut poser l'hypothèse — sans préjuger de sa pertinence psychologique ou neurolinguistique — de l'existence de deux choix, successifs ou non, du locuteur :

1) choix des lexies, étroitement dicté par le contenu sémantique à exprimer. Si choix il y a, il se limite, à un certain niveau, à la synonymie. L'élément de choix serait donc particulièrement étroit en terminologie, si l'on devait se ranger à l'opinion répandue, mais à notre avis bien contestable, concernant la quasi-inexistence de synonymes en terminologie ;

2) choix des structures syntaxiques. Il s'agit du choix de l'organisation, à l'intérieur d'une structure, des lexies sélectionnées. Ce choix est beaucoup plus ouvert (cf. Mel'čuk selon qui, pour une phrase donnée il peut y avoir plusieurs centaines de milliers de paraphrases possibles : communication personnelle, Montréal, mai 1986). L'ouverture même du choix, et son corollaire de faible prévisibilité, fait de la phrase une unité difficile à intégrer dans un recueil d'encodage. C'est pourquoi on doit tendre vers la mise en évidence de groupements dont l'amplitude sera réduite selon des critères qui seront étudiés ci-dessous. Mais, même parmi ces groupements, certains des éléments entretiennent des rapports de statuts très variés qu'on peut résumer ainsi :
   - ◆ groupements dont le sens global ne peut pas être dérivé de la somme des sens des éléments. Il s'agit des idiotismes déjà évoqués, du genre *kick the bucket* ;
   - ◆ groupements dont le sens global peut être partiellement dérivé de la somme des sens des éléments, ex. : *sabler le champagne* ;
   - ◆ groupements dont le sens global peut être totalement dérivé de la somme des sens des éléments, ex. : *mesurer une température*. (Cette classification est très semblable à celle de Cowie, qui distingue : «Pure idioms», «Figurative idioms», «Restricted collocations» et «Open collocations»; 1983 : xii-xiii).

On observera que dans les idiotismes, la suppression d'un des éléments laisse l'autre «en déséquilibre». Pour les quasi-idiotismes comme *sabler le champagne*, l'un des éléments (*champagne*) conserve son sens lexical ordinaire (ce qui n'était le cas ni pour *kick* ni pour *bucket*). Enfin, dans le troisième type de groupements, chacun des éléments conserve son sens lexical primitif. Il existe donc une sorte de dissymétrie entre les deux (ou plus) éléments des quasi-idiotismes qu'on ne retrouve ni dans les idiotismes purs, ni dans les groupements du troisième type. Ce sont, parmi les groupements symétriques, ceux du troisième type (ex. : *mesurer une température*) qui nous intéressent.

II. CRITÈRES QUANTITATIFS

Cependant, tous ces groupements ne sont pas également intéressants (cf. référence à la notion de fréquence dans la définition ci-dessus et le point de vue de Greenbaum — 1974 : 82 — «Virtually any two items can cooccur at a given arbitrary distance»). Deux problèmes distincts, bien que liés, sont à considérer au plan quantitatif, en discours :

- ■ la distance des éléments dans le groupement ;
- ■ la fréquence des groupements eux-mêmes, considérés par rapport à la fréquence de chacun des éléments.

Le premier point concerne ce qu'on peut appeler «empan» (traduction proposée pour l'anglais «span», «A "span" is the co-text within which the collocates are said to co-occur» : Martin, Al et Van Sterkenburg 1983 : 84). Il est clair que les éléments qui ont tendance à fonctionner ensemble se trouvent d'autant plus proches dans le texte que cette tendance est forte. Il y a des exceptions, mais elles ont souvent une valeur esthétique (cf. les incises, les inversions...), parfois rhétorique (cf. l'organisation des périodes, par exemple) et sont donc moins fréquentes dans les textes techno-scientifiques que dans les textes littéraires.

Il s'agit donc de déterminer la taille de l'empan et le point à partir duquel cet empan est appliqué (Berry-Rogghe 1973 : 105). Un exemple simple peut éclairer le problème. À partir de «L'imperfection des sens de l'observateur est responsable de la dispersion des résultats», un des points de départ (ou nœuds) possibles est «imperfection». Si l'empan est égal à 5, on obtient le groupement «imperfection — observateur», si l'empan est égal à 2 on obtient «imperfection — sens». On doit donc travailler avec deux variables, et deux types de solutions peuvent être envisagés :

■ *un travail exhaustif*

Dans ce cas, chaque mot peut servir successivement de nœud, et l'empan prend toutes les valeurs comprises entre 1 et une limite à fixer, mais qui peut être grande (observer que la fixation de cette limite est tout de même en partie arbitraire).

■ *un travail sélectif*

Dans ce cas, seuls certains mots sont pris comme nœuds, selon des critères à déterminer qui peuvent être sémantiques, grammaticaux, statistiques (voir, par exemple, Geffroy *et al.* 1973 : 114, pour un recours au critère de fréquence), etc. L'empan aura alors une valeur fixe jugée optimale (par ex. 5 ou 6 ; elle est de cinq mots à gauche et cinq à droite pour Martin, Al et Van Sterkenburg 1963 : 84). On évite ainsi les suites trop longues dont l'exploitation ultérieure est difficile (voir, par exemple, les possibilités de recouvrement de deux groupements à l'intérieur d'un même fragment de texte) ou les suites trop brèves qui peuvent interdire le repérage de groupements intéressants mais de taille supérieure à l'empan.

On peut observer que les deux types de méthode (exhaustif *vs* sélectif) constituent deux pôles d'un continuum et que les travaux quantitatifs sur la cooccurrence se situent entre ces deux extrêmes, plus ou moins près de l'un ou de l'autre. Il est, en tout cas, indispensable de prendre une décision concernant à la fois le choix des nœuds, et celui de la valeur de l'empan.

Dans le cadre d'une étude préliminaire, un travail exhaustif n'est pas envisageable. Pour un travail sélectif, le premier choix concerne les nœuds. Parmi les critères possibles, nous en examinerons seulement trois, qui peuvent faire l'objet de combinaisons :

1) Critères de fréquence

On peut estimer que les termes les plus intéressants sont les plus fréquents. Le problème de la représentativité du texte sur lequel on travaille est évidemment important, mais on le supposera résolu soit par le soin apporté à la sélection elle-même, soit par la possibilité d'exercer des compensations par la prise en compte de textes supplémentaires. On est donc dans le cadre connu, bien que recélant des problèmes délicats, de la composition d'un corpus représentatif (cf. pour la langue commune, en anglais le Brown Corpus — KUCERA et FRANCIS 1967 — ou le LOB Corpus — JOHANSON *et al.* 1978 — ou, en français, le corpus du TLF — IMBS 1971).

S'agissant de groupements, on est confronté à un problème supplémentaire. Puisque l'aspect intéressant concerne non pas tel ou tel élément du groupement mais la relation de cooccurrence des éléments, il est clair que le critère de fréquence portant sur une partie du groupement seulement n'est pas satisfaisant. On peut, en effet, imaginer deux éléments ayant chacun une fréquence moyenne ou faible dans le texte mais dont la fréquence de cooccurrence serait élevée par rapport aux autres fréquences de cooccurrence. Il faut donc être très attentif lorsqu'on veut recourir à ce critère quantitatif pour la sélection des nœuds.

2) Critères sémantiques

Dans le cadre d'une étude très fine, on peut avoir à travailler sur un domaine sémantique ou terminologique précis. Il est souhaitable alors de prendre comme nœuds des mots faisant partie de la langue spécialisée du domaine. On retrouve alors les pratiques lexicographiques traditionnelles de collecte des contextes en vue de la rédaction d'articles de dictionnaires.

3) Critères grammaticaux

On peut s'intéresser aux groupements contenant des éléments appartenant à telle ou telle catégorie grammaticale. C'est une pratique particulièrement féconde pour un travail sur les mots-outils. Bien qu'on ait exclu ce type de mots du champ de la présente recherche, l'intérêt du critère grammatical subsiste. On sait que la terminologie est un grande pourvoyeuse de formes nominales (96 % dans Baudot et Clas 1984 : 49 ; 94 % dans Clas et Baudot 1985 : 5) et il peut être intéressant, pour toutes sortes de raisons, d'examiner les différents groupements qui peuvent se constituer autour de nœuds nominaux : on peut être ainsi conduit à des observations de type grammatical et sémantique intéressantes pour le chercheur comme pour le traducteur. Si la forte densité nominale d'un texte technique est un argument en faveur de l'étude des groupements à nœud nominal, c'est aussi un motif de prudence lors d'un choix concernant un travail exploratoire comme le nôtre. Nous avons donc préféré nous intéresser à une autre catégorie, importante elle aussi mais beaucoup moins nombreuse (Lafon 1984 : 193), celle du verbe.

Parmi les groupements contenant des verbes, il faut, en outre, prévoir les conditions syntaxiques au phénomène de cooccurrence. C'est d'ailleurs ce que propose Greenbaum : «Collocability should be tied to syntax, though a syntax that caters to connections between sentences» (1974 : 82). C'est aussi ce que font Benson *et al.* dans leur *Combinatory Dictionary*. Dans le cas présent, à partir du verbe par exemple, on dédaignera les groupements verbe + préposition pour se concentrer sur les groupements verbe + nominal ou verbe + adverbial (à l'exclusion des groupements verbe + adverbe qui constituent une seule lexie, assimilables aux idiotismes).

L'application des critères ci-dessus conduit à sélectionner une catégorie spéciale de groupements préférentiels, qui se caractérise premièrement par le fait que chacun de leurs éléments garde son sens individuel, et, deuxièmement, par le fait que la fréquence du groupement est plus élevée que ne le laisse prévoir la fréquence des éléments qui le constituent. On les appellera *collocations* et leurs éléments constitutifs seront nommés *cooccurrents*, sans qu'on cherche à établir entre eux une hiérarchisation du type base *vs* collocatif (Hausmann 1979 : 189).

C'est donc un recueil de collocations en langue techno-scientifique sectorielle qu'il s'agit de compiler. Un tel recueil est-il possible ? C'est la question que Hausmann posait en 1979, mais à propos de la langue commune.

Pour Hausmann, la rédaction d'un dictionnaire de collocations pose des problèmes pour deux raisons.

- Il est, selon lui, «impossible de séparer, en théorie, les combinaisons acceptables (type *longs cheveux*) des combinaisons non acceptables (type *grands cheveux*) car il est souvent possible d'inventer un contexte qui rende acceptable une combinaison inattendue» (Hausmann 1979 : 190).
- En outre, «le nombre de combinaisons acceptables dans une langue donnée est si élevé qu'il serait vain de vouloir les réunir toutes dans un dictionnaire» (*ibid.*).

Lorsqu'il remplace «acceptable» par «probable» ou «usuel», Hausmann constate que sa première objection demeure : en effet, pour lui, la probabilité d'une collocation ne pourrait être estimée que grâce à sa fréquence dans un corpus immense, qu'il pense — sans donner d'explications sur cet ordre de grandeur — devoir être «mille fois plus important que celui qui a servi à la réalisation du *Trésor de la langue française* (Hausmann 1979 : 191), soit environ 70 milliards d'occurrences. D'autre part, Hausmann considère que «le nombre des combinaisons probables est toujours très élevé» (*ibid.*).

Concernant la première objection, il faut savoir que la validité de la notion de probabilité en langue est discutée. Depuis les travaux de Müller (1964 : 114 et suiv.), on a beaucoup réfléchi à cette question et on a pris conscience du rôle important joué par les

facteurs thématiques et stylistiques. Contrairement à ce que suggère Hausmann, il n'est pas certain que le seul accroissement de la taille d'un corpus permette d'estimer convenablement la *probabilité en langue* d'un vocable. Lafon fait valoir, par ailleurs, que, s'agissant non plus d'éléments simples et nombreux, comme le sont les phonèmes par exemple, mais de rencontres entre des vocables dont la fréquence est parfois faible, aucun modèle statistique simple ne permet d'estimer la probabilité en langue (Lafon *et al.* 1985 : 60). Il semble donc que, même s'il ne s'appuie pas sur des considérations statistiques, Hausmann a raison d'être pessimiste en ce qui concerne son premier point.

En revanche, il y a confusion sur le second. En toute rigueur, on ne peut pas dire, dans un premier temps, qu'il est impossible de distinguer entre ce qui est probable et ce qui ne l'est pas pour faire valoir ensuite que le nombre des combinaisons *probables* est très élevé. Il y a là un glissement dans l'emploi de l'adjectif «probable», dont le sémantisme recèle en effet bien des pièges.

En statistique linguistique, on ne peut pas parler de «combinaisons probables» dans l'absolu. Il faut évidemment préciser le seuil où l'on se place. C'est la valeur de ce seuil qui est d'ailleurs intéressante dans la perspective lexicographique adoptée par Hausmann. On voit bien, en effet, que si l'on souhaite limiter, en fonction de critères statistiques, le nombre de collocations à intégrer dans un dictionnaire, il suffit de faire varier le seuil. On aura ainsi une liste de collocations dont la probabilité est supérieure à 0,01, une autre où cette probabilité est supérieure à 0,05, etc. Tout ceci, répétons-le, dans le cas où l'on jugerait acceptable l'estimation d'une probabilité en langue à partir du cadre d'un corpus.

C'est bien là qu'il faut situer le vrai problème. Il existe toute une tradition lexicographique qui se fonde sur la lecture de documents divers assortie d'une évaluation intuitive des fréquences «significatives». En d'autres termes, il est plus d'un lexicographe qui, ayant relevé dans son corpus un nombre qu'il estime suffisamment grand d'occurrences d'une même forme, décide d'inclure cette forme dans son dictionnaire. Qu'il considère ensuite avoir dressé un dictionnaire de *langue* est une autre affaire dont on ne dira rien ici (cf. Rey-Debove 1971 : 90).

Hausmann par exemple, mais aussi Rodale (1947), Reum (1955), Katsumata (1958) ou Friederich et Canavan (1979), se placent dans la perspective de l'étude de la langue commune. Dans le cadre du travail exploratoire où nous nous situons, le problème de l'homogénéité du corpus, et donc de sa représentativité avec tout son cortège de conséquences, y compris statistiques, est moins aigu. Il serait naïf de prétendre qu'à l'intérieur d'un domaine techno-scientifique donné, la langue présente une uniformité sans commune mesure avec celle de la langue dite commune. Il existe là aussi des différences de style d'un texte à l'autre, même si les caractéristiques de la langue techno-scientifique ne sont pas celles de la langue commune (cf. Crystal et Davy 1969). Toutefois, pour les questions qui nous concernent ici et qui sont surtout d'ordre lexico-sémantique, l'homogénéité est supérieure. Ceci doit donc entraîner des fréquences d'emploi de vocables plus élevées elles aussi et, ce faisant, des fréquences de collocations plus facilement utilisables.

La question de l'inclusion de telle collocation dans le dictionnaire se pose surtout, pour des raisons évidentes de coût, de maniabilité, etc., pour le dictionnaire imprimé. Or, on parle de plus en plus de la diffusion des ouvrages de référence, et donc des dictionnaires, sur des supports qui les rendent lisibles par ordinateur (Baudot 1986 : 153-158). S'agissant d'ouvrages de grande diffusion concernant la langue commune, la question peut effectivement prêter à controverse. Mais dans le cadre d'un ouvrage de référence techno-scientifique, où les problèmes de mise à jour rapide sont primordiaux, il est probable que la diffusion d'outils documentaires informatisés se généralisera auprès d'un public de professionnels qui de toute façon utilise de plus en plus l'ordinateur (cf. Van Deth 1985).

Cette modification des habitudes de travail ne peut pas ne pas avoir d'incidence sur la nature des outils et, partant, sur la réflexion méthodologique préalable à leur conception et à leur confection. Il faut, en particulier, tenir compte du fait que les modalités de consultation d'un ouvrage de référence informatisé sont sensiblement différentes de celles d'un ouvrage imprimé. On tend à l'utiliser comme une base de données dans laquelle certaines informations doivent être accessibles très rapidement et très facilement, alors que d'autres, supposées moins fréquemment recherchées ne seront obtenues que plus lentement. Les liaisons entre les éléments d'un tel ouvrage doivent être établies par le système informatique, afin que l'utilisateur soit en mesure de les retrouver simplement.

Dans le cadre d'un index de collocations, l'accès doit être obtenu par l'un ou l'autre des cooccurrents. Chacun de ces éléments doit être relié, en réseau, à d'autres ensembles. Un tel système risque, si l'on n'y prend garde, de submerger l'utilisateur sous un flot d'informations, obtenues d'ailleurs au prix d'un temps de consultation d'autant plus long que l'ouvrage sera plus complet. C'est pourquoi il faut hiérarchiser les données et structurer cet outil. Si l'on admet, pour commencer (mais voir ci-dessous), et compte tenu de la relative homogénéité du corpus, que la fréquence d'emploi d'une collocation reflète convenablement son intérêt pour le traducteur/rédacteur, on pourra utiliser cet élément comme critère fondamental de hiérarchisation des données. C'est aux collocations les plus fréquentes qu'on aura le plus rapidement accès.

Ainsi, à partir de LOI on aboutira facilement à ABROGER, VOTER, ..., RESPECTER, VIOLER, ..., mais aussi, moins rapidement peut-être, à AIMER, DÉTESTER, ... L'intérêt de ces derniers cooccurrents n'est pas nul, s'agissant répétons-le d'un ouvrage d'encodage, puisqu'il permet au traducteur, sinon au rédacteur, d'éviter des verbes comme ADORER, ABHORRER, ..., qui, sans être impossibles avec LOI, donneraient au texte une originalité qui pourrait être prise en mauvaise part.

Une bonne partie des problèmes serait réglée si l'on était sûr que le critère de fréquence d'emploi dans un corpus coïncide avec celui d'utilité. Observons d'abord que ce dernier critère est bien vague et susceptible de fluctuer avec les individus mais aussi avec le temps. Tel événement de portée nationale ou internationale peut faire apparaître un vide terminologique pour lequel tel groupement jusqu'alors peu usité se révélera bien utile. Il existe néanmoins, au-delà de ces variations individuelles ou circonstancielles, un fonds commun dont les spécialistes du domaine savent bien qu'il est indispensable. Même si l'on admet qu'il puisse y avoir quelques surprises — comparables à celles qui furent révélées lors du travail sur le français fondamental et qui ont conduit au concept de *disponibilité* — on peut estimer que, s'agissant de langue *écrite* dans le cadre d'un domaine techno-scientifique bien délimité, le critère de fréquence (auquel on pourrait le cas échéant songer à adjoindre celui de la répartition) est suffisamment opérationnel pour servir de base de départ.

Il faut en effet envisager dès le début le caractère évolutif de ces index de cooccurrents. Comme déjà dit, la nature même de la langue techno-scientifique implique la prise en compte de l'évolution de son stock lexical. Mais on peut aussi prévoir des ajustements de la structure des données. Rien ne doit être figé dans un outil de ce genre, et il faut laisser toute latitude au concepteur d'abord, voire à l'usager ensuite, d'adapter l'ouvrage à ses besoins. Ceci suppose, entre autres, que la probabilité de base de telle collocation puisse être modifiée afin que le système réagisse de manière plus «personnalisée» selon les besoins de l'utilisateur.

On en arrive ainsi au concept de «version standard» de l'index, susceptible d'être remise à jour périodiquement par le concepteur mais capable d'intégrer, outre ces mises à jour, des modifications répondant aux nécessités spécifiques de chaque utilisateur. Il faut signaler ici que ce type de solution est d'ores et déjà adopté dans le domaine de la

traduction technique où le traducteur se constitue, en fonction de la source de son texte, un mini-glossaire spécifique permettant de prendre en compte les particularités des différents fabricants par exemple.

En concevant un index de cooccurrents en langue techno-scientifique sectorielle comme un ensemble à la fois évolutif et cumulatif — grâce aux possibilités de l'informatique — on parvient à se libérer des objections parfois formulées à l'encontre de l'élaboration des dictionnaires de collocations. Le problème peut être posé en termes radicalement différents. Il ne s'agit plus de trancher de manière définitive quant à l'inclusion d'un groupement. Si le groupement a été relevé, il doit figurer dans l'index avec sa fréquence d'emploi, fût-elle de 1 seulement. Les modalités des opérations de relevé doivent être précisées, mais elles se fondent ici sur une analyse des contextes droits et gauches des bases verbales avec un empan allant de 1 à 5 ou 6. Si le corpus est lisible par ordinateur, ces opérations de relevé peuvent être faites assez rapidement, quitte à procéder ultérieurement à des opérations de remise en ordre syntaxique (c'est-à-dire «des mesures seront prises» avec «on prendra des mesures») ou de stricte lemmatisation.

On doit aboutir à un index structuré sur des bases aussi bien statistiques que terminologiques dont les clés d'accès seront les cooccurrents. Ainsi, en interrogeant par LOI on devra obtenir :

- la liste des n (par exemple 20) collocations les plus fréquentes incluant LOI ;
- la possibilité de renvoyer à toutes les autres collocations présentes dans le corpus ;
- la possibilité de consulter la liste des éléments de même catégorie grammaticale (c'est-à-dire ici «nominal») ayant un rapport terminologique avec LOI (c'est-à-dire DÉCRET, ARRÊTÉ,...).

Il peut se révéler nécessaire, lorsque l'index atteint plusieurs dizaines de milliers de cooccurrents, de distinguer en son sein entre plusieurs fichiers de collocations. Un premier fichier serait destiné à la consultation rapide des collocations fréquentes, un second fichier à la consultation «lente» des collocations plus rares, et enfin un troisième fichier, ayant le statut de répertoire de travail, pour des collocations de fréquence 1. Au fur et à mesure de l'évolution du corpus, le contenu de chacun des fichiers devra, bien sûr, subir des modifications qui affecteront la fréquence de telle collocation et pourront, éventuellement, la faire passer d'un fichier à un autre.

Si nous parlons d'index *cumulatif*, c'est pour prendre en compte la possibilité et la nécessité d'accroître le corpus afin d'intégrer les nouveautés lexicales et collocationnelles dont on a dit qu'elles peuvent être nombreuses en langue techno-scientifique. Si nous parlons d'index *évolutif*, c'est, bien sûr, parce que cette caractéristique est inévitablement liée à la précédente, mais aussi parce que c'est bien le caractère dynamique de ce genre d'index qui en fait l'intérêt.

Ce type d'outil de référence est donc en définitive assez éloigné du dictionnaire classique. Sa conception même, l'utilisation qui peut en être faite, son caractère évolutif le rapprochent plutôt des systèmes de gestion de bases de données auxquels il empruntera nécessairement pour la solution informatique des problèmes à résoudre. C'est cependant au linguiste terminologue qu'il appartient de rassembler le corpus et d'en extraire les groupements à inclure ; c'est à l'utilisateur, rédacteur ou traducteur, qu'il appartiendra de tester l'outil et de le modifier éventuellement, pour son usage personnel ou pour une meilleure performance générale. Il est clair que la qualité du travail fourni par ces deux parties sera déterminante, même si l'informaticien est la cheville ouvrière de l'entreprise.

NOTE

Ceci suppose une conception de la synonymie identique à celle de Mel'čuk, pour qui deux mots sont synonymes *salva significatione* même si leurs restrictions d'emploi

sont différentes. Mel'čuk cite l'exemple archi-connu de *gravement* et *grièvement* («gravement malade» et «grièvement blessé», *vs* «*grièvement malade» et «?gravement blessé»). Pour lui, ces deux adverbes sont synonymes.

BIBLIOGRAPHIE

BALLY, Charles (1951, 3e éd.) : *Traité de stylistique française*, Paris, Klincksieck.

BAUDOT, Jean (1986) : «Les banques de terminologie de l'avenir», *META*, 31:2, pp. 153-158.

BAUDOT, Jean et André CLAS (1984) : «A Model for a Bilingual Terminology Mini-Bank», *Lebende Sprachen* 2, pp. 49-54.

BENSON, Morton (1985) : «A Combinatory Dictionary of English», *Dictionaries* 7, pp. 189-200.

BENSON, Morton, Evelyn BENSON and Robert ILSON (1985) : *The BBI Combinatory Dictionary : A Guide to Word Combinations in English*, Amsterdam, Benjamins.

BERRY-ROGGHE, Godelieve L.M. (1973) : «The Computation of Collocations and their Relevance in Lexical Studies», in AITKEN, A.J. *et al.*, eds. : *The Computer and Literary Studies*, Edinburgh University Press, pp. 103-112.

CLAS, André et Jean BAUDOT (1985) : «BATEM. Une banque de terminologie sur micro-ordinateur», communication au 2e Symposium d'Infoterm, «Travail dans le cadre d'un réseau de terminologie», Vienne, avril 1985.

COWIE, Anthony P. (1981) : «The Treatment of Collocations and Idioms in Learners' Dictionaries», *Applied Linguistics* 3, pp. 223-235.

COWIE, Anthony P., Ronald MACKIN and I.R. McCAIG (1983) : *Oxford Dictionary of Current Idiomatic English*. Vol. 2 : *Phrase, Clause and Sentence Idioms*, Londres, Oxford University Press.

CRYSTAL, David and Derek DAVY (1969) : *Investigating English Style*, Londres, Longman.

DZIERZANOWSKA, Halina and Christian DOUGLAS KOZŁOWSKA (1982) : *Selected English Collocations*, Varsovie, Pánstwowe Wydawnistwa Naukowe.

FALK, J.S. (1973) : *Linguistics and Language*, Lexington.

FIRTH, J.R. (1957) : *Papers in Linguistics* 1934-1951, Londres, Oxford University Press.

FRIEDERICH, Wolf and John CANAVAN (1979) : *Dictionary of English Words in Context*, Dortmund, Lensing.

GAK, B.G. (1963) : *Dictionnaire de phraséologie française*, Moscou.

GEFFROY, Annie, P. LAFON, Gill SEIDEL, M. TOURNIER (1973) : «Lexicometric Analysis of Co-occurrences» in AITKEN, A.J. *et al.*, eds. : *The Computer and Literary Studies*, Edinburgh University Press, pp. 113-133.

GREENBAUM, Sidney (1974) : «Some Verb-Intensifier Collocations in American and British English», *American Speech* 49, 1-2, pp. 79-89.

HALLIDAY, M.A.K. (1986) : «Lexis as a Linguistic Level», in BAZELL, C.E. *et al.*, eds. : *In Memory of J.R. Firth*, Londres, Longman, pp. 148-162.

HAUSMANN, Franz J. (1979) : «Un dictionnaire de collocations est-il possible ?», *TraLiLi*, 17, 1, pp. 187-195.

HOCKETT, Charles F. (1958) : *A Course in Modern Linguistics*, New York, MacMillan.

IMBS, Paul (1971) : «Préface» in *Trésor de la langue française*, vol. 1, Paris, CNRS.

JOHANSSON, S., G.N. LEECH, H. GOODLUCK (1978) : *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*, University of Oslo.

KATSUMATA, Senkichiro (1958) : *Kenkyushas New Dictionary of English Collocations*, Tokyo, Kenkyusha.

KUCERA, H. and W. Nelson FRANCIS (1967) : *Computational Analysis of Present Day American English*, Providence, Rhode Island, Brown University Press.

LACROIX, Ulysse (1947) : *Dictionnaire des mots et des idées*, Paris, Nathan.

LAFON, Pierre (1984) : *Dépouillements et statistiques en lexicométrie*, Genève, Slatkine.

LAFON, Pierre, A. SALEM et M. TOURNIER (1985) : «Lexicométrie et associations syntagmatiques», in CHARPENTIER, C. et J. DAVID : *la Recherche française par ordinateur en langue et littérature*, Genève-Paris, Slatkine-Champion, pp. 59-72.

LEECH, Geoffrey (1981, 2e éd.) : *Semantics*, Harmondsworth, Pelican.

LEES, Robert L. (1960) : *A Grammar of English Nominalizations*, Indiana University Research Center in Anthropology, Folklore and Linguistics, Publication 12.

MACKIN, Ronald (1978) : «On Collocations : «Words Shall Be Known by the Company They Keep»», *in* STREVENS, Peter, ed., *In Honour of A.S. Hornby*, Londres, Oxford University Press, pp. 149-165.

MARTIN, W.J.R., B.P.F. AL and P.J.G. VAN STERKENBURG (1983) : «On the Processing of a Text Corpus», *in* HARTMANN, R.R.K. ed. : *Lexicography : Principles and Practice*, Londres, Academic Press, pp. 77-87.

MEIER, Hans H. (1975) : «On Placing English Idioms in Lexis and Grammar», *English Studies*, 56, 33, pp. 231-244.

MULLER, Charles (1964) : *Essai de statistique lexicale : l'Illusion comique de Pierre Corneille*, Paris, Klincksieck.
REUM, Albrecht (1955) : *A Dictionary of English Style*, Leverkusen, Gottschalksche Verlagsbuchhandlung.
REY-DEBOVE, Josette (1971) : *Étude linguistique et sémiotique des dictionnaires français contemporains*, La Haye, Mouton.
ROBINS, R.H. (1964) : *General Linguistics : An Introductory Survey*, Londres, Longmans.
RODALE, J.I. (1947) : *The Word Finder*, Emmaus, Pa., Rodale Press.
SAUVAGEOT, Aurélien (1964) : *Portrait du vocabulaire français*, Paris, Larousse.
SINCLAIR, J. Mch. (1966) : «Beginning the Study of Lexis», in BAZELL, C.E. *et al.*, eds., *In Memory of J.R. Firth*, Londres, Longman, pp. 410-430.
SWEET, Henry (1889) : *The Practical Study of Languages*, Londres, Oxford University Press.
VAN DETH, Jean-Pierre (1985) : *la Traduction et l'interprétation en France*, Paris.
VENDLER, Zeno (1968) : *Adjectives and Nominalizations*, La Haye, Mouton.