Sur les caractéristiques de la collocation

Geoffrey Williams

CRELLIC – Université de Bretagne Sud
Lorient
Geoffrey Williams@univ-ubs.fr

Résumé – Abstract

Le terme " collocation " a été introduit dans les années '30 par J. R. Firth, membre-fondateur de l'école contextualiste britannique, pour caractériser certains phénomènes linguistiques de cooccurrence. Ce phénomène est maintenant accepté comme central dans la compétence linguistique des locuteurs natifs et de grande importance pour l'enseignement, la traduction, la lexicographie, et dorénavant, le TALN. Malheureusement, le concept est difficile à formaliser et ne peut être étudié que par rapport à des exemples prototypiques. Quatre caractéristiques sont analysées, leur nature habituelle, lexicalement transparente, arbitraire et syntactiquement bien formée. Les avantages et inconvénients de chaque critère sont discutés.

The term collocation was first introduced in the thirties by J.R. Firth, founder of the British Contextualist school of thought in linguistics, to characterise certain forms of lexical co-occurrence. Gradually this phenomenon has been seen as a central element in native speaker competence and of great importance in teaching, translation,lexicography, and, now, Natural Language Processing. Unfortunately, the concept is difficult to tie down and can only be studied through reference to prototypical examples. Four characteristics of collocation are discussed their habitual, arbitrary, lexically transparent and grammatically well-formed nature. The advantages and drawbacks of each are considered.

1 La notion de collocation

1.1 Introduction

Le terme " collocation " a été introduit dans les années '30 par J. R. Firth, membre-fondateur de l'école contextualiste britannique, pour caractériser certains phénomènes linguistiques de cooccurrence qui relèvent essentiellement de la compétence linguistique des locuteurs natifs (Firth 1957). De par sa nature, la collocation demeure un concept difficilement formalisé, aucune définition ne satisfait tout le monde. De ce fait les grammairiens et les sémanticiens traditionnels ont tendance à l'ignorer, l'exception étant quelques sémanticiens lexicalistes comme Cruse (1986). L'étude de la collocation est surtout une tâche pratique qui vise à aider

les apprenants, de tout niveau linguistique, et les traducteurs. Ceci a pour effet que le phénomène ne peut être décrit qu'en termes de prototypes, les définitions peuvent citer des cas typiques comme *effreindre la loi* - *break the law*, *buveur invétéré* - *heavy drinker* et accepter qu'il y ait des cas plus ou moins acceptables suivant les applications. Cependant, le fait qu'il existe un continuum entre collocation, idiome, locution et termes n'implique pas que tout soit possible.

Les linguistes ont adopté des critères pour définir ce qui peut être accepté comme collocation. Les théories et approches différents de ce phénomène ont tendance à privilégier certains aspects plus que d'autres, tout en reconnaissant cette notion de centralité.

Les collocations sont des liens syntagmatiques :

- habituels
- lexicalement transparents
- arbitraire
- syntactiquement bien formés.

Nous pouvons regarder chacun de ces critères afin de vérifier leur fiabilité.

1.2 Habituel

La notion de co-occurrence habituelle est le premier critère de Firth - "Collocations of a given word are statements of the habitual or customary places of that word in collocational order" (Firth 1957).

Qu'est ce que cela veut dire exactement ? N'importe quel anglais reconnaîtra facilement l'exemple de Firth "*silly ass*" - *imbécile* comme collocation, même si cette formule est plutôt veillotte. La plupart des entrées du dictionnaire des collocations de Benson et al. (1986) seront reconnues de la même manière, tout en acceptant des variantes régionales ou spécialisées. Comme dans le cas des exemples de Firth, ces entrées sont basées sur l'intuition d'un lexicographe professionnel et seront immédiatement reconnues par des locuteurs natifs. Bien que nous considérions normal de rencontrer de tels binômes, nous n'avons toujours pas défini ce que nous voulons dire par 'habituel'. Dans le Petit Robert nous trouvons "*Qui tient de l'habitude par sa régularité, sa constance"*, ceci ne nous avance guère, une telle définition n'etant pas utilisable ni pour une définition linguistique, ni pour une extraction informatique. Pouvons-nous améliorer notre compréhension d'habituel en appliquant une mesure statistique ?

La réponse de l'école de la linguistique de corpus de Birmingham est que :

> SIGNIFICANT COLLOCATION is regular collocation between two items, such that they co-occur more often than their respective frequencies and the length of the text in which they appear would predict. (Sinclair 1970 : 150.)

Ceci exclut effectivement les simples cooccurrences de hasard, mais laisse ouvert le choix de la mesure statistique la plus appropriée. Les expérimentations ont nombreuses et incluent le z-score, le t-score et l'information mutuelle; Daille (1994) donne une liste longue, mais non exhaustive des méthodes de mesures adoptées par tel ou tel chercheur. Chaque outil statistique semble approprié pour une application donnée, ce qui impliquerait que la définition de la collocation adoptée pour l'application est fonction de la mesure statistique appliquée. Cette situation est bien résumée par Clear (1993: 282) pour le t-score et pour l'information mutuelle quand il note que :

> The t-score statistic, by identifying frequent and very reliable collocations, offers the lexicographer a semantic profile of the node word and a set of particular fixed phrases, grammatical frames and typical stereotypical combinations…

tandis que

> The MI is best consulted for information about pairs which, though not likely to be typical of the usage of the node word, will be strongly associated and tend to form idioms, proverbs and technical phrases

Autrement dit, même la définition statistique de la collocation dépend de l'application, nous pouvons avoir la collocation générale avec l'une et la collocation technique avec l'autre. Le problème sera de décider où passe la limite entre l'utilisation technique et non technique. Nous pouvons être tentés de dire que les mesures statistiques ne fournissent qu'une collocation "candidat" et que seul le lexicographe, traducteur ou linguiste natif peut décider ce qui constitue une collocation véritable. Une autre difficulté à résoudre sera de décider où nous placons la limite entre véritable collocation et cooccurrence banale.

Ceci nous mène à une difficulté évoquée par Haussman (1985) pour lequel il y a un continuum entre des collocations libres ou banales, et donc relativement sans intérêt et les collocations figées. Où fixer la limite entre ce qui est significatif et ce qui est banal ? Dans un texte plus récent, Haussman (1997) a même déclaré que "tout est idiomatique". Prenons le cas des maladies et le verbe avoir. Ceci forme des collocations banales comme *il a eu un rhume*, *il a eu un infarctus etc.,* ce qui correspond à *he had a cold*, *a heart attack* etc., en anglais. Ceci semble relativement banal, mais il faut noter deux choses ; la première est que nous parlons des infarctus quasi uniquement au passé pour les deux langues, et que pour le Français nous pouvons également le traduire par *il a fait un infarctus*, par contre *he made a heart attack*, n'est pas possible en anglais.

Il est évident que tandis que la notion d'habituel est importante comme critère, c'est un critère qui reste flou. Si nous regardons le langage humain comme un processus, et pas simplement un produit, ce qui est habituel va changer au cours du temps et selon les facteurs régionaux, techniques et niveau social. La collocation doit être conçue comme phénomène dynamique dont la signification est négociée et non figée. La notion du figement nous posera également des difficultés avec le prochain critère, la transparence lexicale.

1.3 Lexicalement transparent

Dans la plupart de ses écrits sur la collocation, Sinclair insiste sur le fait que celles-ci sont lexicales:

> Collocation in its purest sense … recognises only the lexical co-occurrence of words (1991: 170)

Cette définition écarte certaines collocations formées d'un mot lexical plus un participe, des formes assez fréquents en anglais. Cependant ceci est sous-entendu dans les exemples de Firth qui a inventé le terme *colligation* pour des collocations grammaticales. Les colligations sont une forme de collocation importante, mais qui ne seront pas traitées ici. Bien que ces colligations forment une classe importante dans l'apprentissage des langues il est impossible de les considérer en termes de transparence. En conséquence je ne considère que les collocations lexicales.

Lorsque nous analysons des collocations, surtout en anglais, nous devons tenir compte du mot orthographique puisque certaines collocations figées se comportent comme des idiomes, ne sont plus décomposables et fonctionnent comme des mots. De ce fait les idiomes comme *colère noire* (Kahane et Polguère 2000) peuvent être écartés du champ des collocations. Une telle condition est imposée par Mel'cuk qui a différencié les phrasèmes (idiomes) et des semi-phrasèmes (collocations) selon leur degré d'opacité sémantique. Cependant, comme le démontrent Moon (1998) et Cermák (2001) il est très difficile d'isoler les deux catégories comme des pôles clairement définis, il y un continuum entre les deux. Ainsi, d'après une analyse statistique, *colère* et *noire* sont une cooccurrence significative, et suivant une analyse en termes de fonctions lexicales *noire* exprime un degré de colère, et peut être décrit avec la fonction *Magn*, qui intensifie. Donc, *noir* n'est pas transparent, ce qui signifie que *colère noire* est un idiome, mais le syntagme peut être traité comme collocation.

La notion de transparence décrite par Cruse n'est pas sans difficulté. Cruse nous donne l'exemple de *heavy drinker*, traduit par buveur invétéré. Dans ce cas, *heavy* ne veut pas dire lourd, mais quantifie le degré, intense, de consommation. Pour que le syntagme soit transparent, il faut accepter que l'adjectif ait deux sens et que c'est ce deuxième sens qui s'applique ici. Le même problème se pose avec *public.* En français nous avons des binômes comme *lieu public, vente publique*, qui signifient ouvert à tout le monde, en anglais nous trouvons *public transport, public baths*, mais *public school* pour une école privée et chère. Dans ce cas il faut traiter certaines formes comme des collocations et d'autres comme des termes figés. Comme *colère noire*, *public school* est habituel, et donc une collocation, bien que la règle de transparence ne soit pas respectée. La difficulté est encore accrue avec des verbes comme faire et avoir qui peuvent donner des formules comme *faire un gâteau*, dans le sens fabriquer, mais aussi *faire l'amour*, où on ne fabrique rien, sauf peut être un enfant. Par ailleurs, certains verbes entrent dans des constructions où le sens est porté par le nom, comme *faire une promenade*, pour *se promener*. C'est la raison pour laquelle Gross (1981) a qualifié ces verbes de verbes support puisqu'ils forment des collocations intéressantes, mais sont souvent sémantiquement vides. Cependant, bien que vide, leur traduction n'est pas nécessairement simple, ce qui nous amène au critère de la nature arbitraire des collocations.

1.4 Arbitraire

Benson (1989:3) nous informe que " collocations should be defined not just as 'recurrent word combinations', but as arbitrary recurrent word combinations" c'est à dire récurrentes, donc habituelles, mais également arbitraires. La nature arbitraire de la collocation nous ramène à Haussman et son continuum entre collocation libre et figée. Dans cette optique, les collocations ne deviennent significatives qu'en termes de traduction, il est nécessaire de savoir

ce qui peut librement être traduit et est donc non-arbitraire et ce qui exige la connaissance collocationnelle. L'arbitraire est alors un critère de traduction qui peut seulement être jugé entre les couples de langues car ce qui est non-significatif dans une combinaison peut l'être dans un autre. Il s'agit donc d' une question de degré, ce qui a fait déclarer à Haussman (1997) que "*tout est idiomatique* ".

La nature arbitraire de la collocation est facilement démontrée avec des combinaisons comme *heavy traffic*, traduit par *circulation intense*. De nombreux exemples peuvent être fournis entre l'anglais, le français et l'allemand, par exemple. Cependant ce critère reste problématique si nous considérons la collocation en termes de compétence du locuteur natif car le locuteur acquiert de nouvelles collocations tout au long de sa vie, sans référence à d'autres langues, bien qu'il ne devienne conscient du fait que ces formules soient des collocations que si on lui demande de les traduire. En outre, si tout est vraiment idiomatique, comment classer ce qui est banal. Si nous retournons à notre exemple des maladies *have a cold*, *avoir un rhume* est banal, mais *have a heart attack* ne l'est pas à cause de la variante *faire un infarctus*.

1.5 Syntactiquement bien formés

Jusqu'ici aucun des critères n'est sans difficulté, la condition de grammaticalité ne l'est pas non plus. La condition est importante pour Kjellmer (1984), mais seulement de son point de vue de lexicographe qui cherche à limiter le bruit dans une extraction automatique de collocations. Dans le cas des collocations lexicales, une solution de facilité serait d'imposer des listes d'exclusions. Les outils plus sophistiqués où les préposition ne peuvent pas être éliminées, comme l'outil Xtract de Smadja (1983), ajoutent des filtres linguistiques afin de nettoyer les collocations et formules.

Les critères grammaticaux peuvent être efficaces dans l'extraction des collocations, mais ils n'offrent pas de réponses absolues. Les collocations textuelles conformes à la définition de cooccurrence dans une fenêtre utilisée par Sinclair n'ont pas besoin de ce critère puisqu'elles recherchent des associations dans un champ sémantique. En effet Firth lui-même a déclaré que :

> *Collocations of a given word are statements of the habitual or customary places of that word in collocational order but not in any other contextual order and emphatically not in any grammatical order.*

Des recherches entreprises dans cette direction comme celles de Berry-Roghe (1973) sur les thèmes, de Phillips (1985) sur la notion de "aboutness", ou thème général, ou celles de Clear (1994) sur la polysémie, ou plus récemment les réseaux collocationels de Williams (1998, 1999) n'utilisent pas l'aspect grammatical

D'autres difficultés affectant le langage naturel en contexte découlent des jugements sur ce qui est acceptable. Comme Sinclair (1984: 203) l'a démontré, beaucoup de phrases bien formées peuvent apparaître anormales pour un locuteur natif. La solution sera, selon Sinclair, de rechercher ce qui semble "naturel" plutôt que ce qui est bien formé.

1.6 Conclusion

Nous avons donc quatre critères pour juger les collocations potentielles et les délimiter par rapport à d'autres locutions et formules idiomatiques. Malheureusement aucun des quatre n'est accepté sans réserve par tous les linguistes, et même quand ils le sont, il reste toujours un certain flou. Ceci nous ramène à la théorie des prototypes où, comme dans l'exemple de jeu de Wittgenstein (1953), nous avons des exemples typiques, centraux, mais tous les membres des classes ne peuvent satisfaire toutes les conditions tout le temps. En effet, en faisant référence à des prototypes nous savons ce que nous voulions dire par collocation, mais ne pouvons formaliser le concept que par référence à une application particulière. Cependant, pour faciliter les choses, nous pouvons isoler deux grandes tendances (Williams 2000) :

- La tendance lexicographique qui tend à une formalisation des collocations pour les inclure dans des dictionnaires.

- La tendance contextualiste, en ligne directe avec les travaux de Firth, qui considère les collocations comme un phénomène textuel et les définit en fonction de l'apparition de cooccurrences à l'intérieur d'une fenêtre.

Le linguiste, le traducteur, le lexicographe, l'informaticien ou l'enseignant peut décider d'accepter ou d'éliminer des critères, d'être plus ou moins strict, mais aucun ne possède toute la vérité sur le phénomène. Les deux tendances décrivent le même phénomène, l'approche est simplement différente.

Références

Benson M. (1989). The Structure of the Collocational Dictionary. *International Journal of Lexicography*. Vol.2 /1 pp 1-14.

Benson M., Benson E., Ilson R. (1986). *The BBI Dictionary of English Word Combinations*. Amsterdam, John Benjamin's

Berry-Roghe G.L.M. (1973). The computation of collocations and their relevance in lexical studies, dans Aitken A.J,. Bailey R., Hamilton-Smith N., (eds*), The Computer and Literary Studies*, Edinburgh, Edinburgh University Press

ČERMAK F. (2001). Substance of idioms: Perennial Problems, Lack of Data or Theory? *International Journal of Lexicography*. vol.14/1. March 2001 pp 1-20.

Clear, J. (1994). I can't see the sense in a large corpus. Papers in Computational Lexicography. Actes de *Complex '99*. Hungary: Budapest. 33-48.

Cruse D.A., (1986*) Lexical Semantics*, Cambridge, Cambridge University Press.

Daille B. (1995). Combined Approach for Terminological Extraction, *Lexical Statistics and Linguistic Filtering*, : Lancaster University Unit for Computer Research on the English Language, Technical Papers 5.

Firth J.R. (1957). A synopsis of linguistic theory 1930-1955 dans Firth J.R., *Papers in Lnguistics 1934-1951*, Oxford, Oxford University Press.

Gross M. (1981) Les bases empiriques de la notion de prédicat sémantique. *Langage*. Vol. 63 pp 7-52.

Haussman F-J. (1985). Kollokationen im deutschen Wörterbuch: Ein Beitrag zur Theorie des lexikographischen Beispiels, dans Bergenholtz H., Mugdan J., (eds) *Lexikographie und Grammatik.* Tübingen, Niemeyer.

Haussman F-J. (1997). Tout est idiomatique dans les langues. dans *La locution entre langue et usages*. Michel MARTINS-BALTAR (éditeur). Paris, ENS Editions. pp 277-290

Kahane S., Polguère A. (2000). Un langage formel d'encodage des fonctions lexicales et son application à la modélisation des collocations. dans Daille B., Williams G. (2000). *Journée d'études de l'ATALA :La Collocation*. Nantes, Rapport de Recherche 00.13 IRIN. Décembre 2000.

Kjellmer G. (1984). Some thoughts on Collocational Distinctiveness. dans Aarts J., Meijs W. (eds) (1984). *Corpus Linguistics: recent advances in the use of computer corpora in English language research*. Amsterdam : Rodopi. pp 163-171.

Mel'cuk I., (1984). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I*, Montréal, Les Presses de l'Université de Montréal.

Moon R. (1998). *Fixed Expressions and Idioms in English.* Oxford, Clarendon Press.

Phillips M. (1985). *Aspects of Text Structure: An investigation of the lexical organisation of text*. Amsterdam, North Holland.

Sinclair J. McH, (1991). *Corpus, Collocation, Collocation*, Oxford, Oxford University Press.

Sinclair J. McH, (1984). Naturalness in Language. dans Aarts J. et MeijsW; (eds) (1984). *Corpus Linguistics: recent advances in the use of computer corpora in English language research.* Amsterdam, Rodopi. pp 203-210.

Sinclair J. McH, et al. (1970). *English Lexical Studies: Report to OSTI on Project C/LP/08.* Birmingham, Department of English, University of Birmingham

Smadja F**.** (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics*. Vol. 19/1 pp 143-177.

Williams G. (2000). Avant propos dans Daille B., Williams G. (2000). *Journée d'études de l'ATALA :La Collocation*. Nantes, Rapport de Recherche 00.13 IRIN. Décembre 2000.

Williams G. (1998). "Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology". *International Journal of Corpus Linguistics*. .3(1): 151-171

Williams G. (1999). *Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique*. Thèse en anglais – linguistique de corpus. Université de Nantes. http://perso.wanadoo/geoffrey.williams

Wittgenstein L. (1953). *Philosophical Investigations*. (Trans. by G.E.M. Anscombe). Oxford, Basil Blackwell (2nd Edition).