Lucie LANGLOIS, Université d'Ottawa
Pierre PLAMONDON, RALI, Université de Montréal

Le repérage automatique de collocations équivalentes à partir de bitextes[1]

Résumé

Rares sont les logiciels conçus spécifiquement pour le repérage automatique de collocations dans un bitexte.[2] C'est pourquoi nous avons décidé d'en créer un qui répondrait aux besoins des lexicographes travaillant au Dictionnaire bilingue canadien. À cette fin, nous avons conçu une méthode qui permet d'obtenir, pour un mot donné, une liste de collocations possibles qui s'articulent autour de ce mot ainsi qu'une liste de collocations équivalentes possibles en langue d'arrivée. Nous expliquerons d'abord notre méthode, puis nous procéderons à une étude de cas.

Mots-clés : Lexicographie bilingue, bitexte, collocation, information mutuelle.

1. Collocation

Les collocations sont un phénomène de langue dont tout dictionnaire, unilingue ou bilingue, doit maintenant tenir compte. Néanmoins, le terme *collocation* est encore assez mal défini. Certains définissent la collocation en termes de "co-occurrence of two or more lexical items as realizations of structural elements within a given syntactic pattern" (Cowie 1978:132) ou de "recurrent word combinations" (Benson *et al.* 1986:vii), d'autres voient la collocation comme étant une "association habituelle d'une unité lexicale avec d'autres unités" (Mounin 1974) ou de "habitual co-occurrence of individual lexical items" (Crystal 1997), tandis que pour d'autres, elle est plutôt une "combinaison phraséologique de deux ou plusieurs mots dans laquelle les mots composants, quoique soumis à une contrainte lexicale, gardent encore leur autonomie de sens" (S.Q. Liang 1991:152). En revanche, si certains définissent encore la collocation en termes de fréquence (Smadja 1993), les nombreux chercheurs sont de plus en plus à se pencher sur l'aspect sémantique du phénomène. Pour Hausmann (1990:1010), par exemple, "la base de la collocation [est] le partenaire caractérisé [...], le collocatif, le partenaire caractérisant qui ne reçoit son identité sémantique que par la collocation".

Et c'est justement à cause de cette dimension sémantique que nous avons pensé d'utiliser un bitexte pour extraire des collocations, le bitexte servant en quelque sorte de filtre *sémantique* permettant de réduire le bruit.

2. Bitexte

L'utilisation des corpus unilingues en lexicographie a fait l'objet de maintes publications, dont la plus connue sans doute est celle qui relate l'expérience de Sinclair et de ses collègues (Sinclair 1987) lors de la rédaction du dictionnaire Collins Cobuild. Cependant, l'utilisation des corpus bilingues en lexicographie bilingue n'a pas encore été beaucoup étudiée, sans doute à cause de la rareté des corpus bilingues et des outils conçus pour les exploiter (McEnery et Wilson 1996:129).

Pour les présents travaux, nous nous sommes servis de huit années du *Hansard* (le Journal des débats du Parlement canadien)[3] apparié selon une méthode à deux temps : l'alignement est d'abord effectué en fonction de la longueur des phrases, puis en fonction des mots apparentés (Simard *et al.* 1992 et Isabelle et Warwick-Armstrong 1993). Ce bitexte est ensuite étiqueté avec les catégories grammaticales, et, pour réduire le bruit, seuls les noms communs (*NomC*), les verbes (*Verb*), les adjectifs (*AdjQ*) et les adverbes (*Adve*) sont conservés. Il semblait raisonnable de ne garder que ces catégories puisque les collocations lexicales (par opposition aux collocations grammaticales, constituées d'un mot dominant suivi d'une unité subordonnée, comme *s'abstenir de* et *absent from*) en français et en anglais s'articulent autour des noms, des verbes, des adjectifs et des adverbes. Finalement, chaque mot du bitexte est lemmatisé, ce qui permet de regrouper sous un même lemme toutes les formes possibles d'un mot, facilitant ainsi le repérage de combinaisons significatives. Les variantes *l'orage éclate*, *l'orage a éclaté*, *l'orage éclatait* et *les orages ont éclaté*, par exemple, apparaissent toutes, dans notre corpus, sous la forme canonique *orage/NomC éclater/Verb.*

3. Techniques utilisées

3.1. Information mutuelle

Dans ce bitexte lemmatisé, nous cherchions à isoler les collocations, qui sont des combinaisons lexicales dont les membres s'attirent l'un l'autre. Ainsi, nous avons choisi le calcul de l'information mutuelle en probabilité puisqu'il permet justement de mesurer le degré de dépendance entre deux événements x et y. L'information mutuelle est définie comme suit:

$$I(x,y) \equiv \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Si les deux événements sont indépendants, $P(x)\ P(y) = P(x,y)$, alors la valeur $I(x,y)$ est 0. Si un événement y est conditionné par x, c'est-à-dire si le fait que x se produise augmente (ou diminue) les chances qu'on observe y, alors $I(x,y)$ sera plus grand (plus petit) que 0.

Church et Hanks (1990) ont utilisé et adapté la notion d'information mutuelle (IM) pour détecter des associations entre des mots d'une langue. Ici les événements x et y sont l'apparition des mots x et y dans une même phrase. $P(x)$ et $P(y)$ sont respectivement la probabilité d'observer un mot x ou y dans une phrase, et $P(x,y)$ est la probabilité d'observer y à l'intérieur d'un empan de k mots suivant x dans un texte (en pratique, k est une constante dont la valeur est fixée à 5 mots). Si effectivement les mots x et y sont associés de quelque façon, la probabilité qu'ils se côtoient, $P(x,y)$, sera plus grande que la probabilité de les observer indépendamment, $P(x)\ P(y)$. Plus les mots sont liés, plus la valeur de l'information mutuelle est grande.

Le calcul de l'information mutuelle en probabilité donne des résultats intéressants pour la détection des collocations à l'intérieur d'une langue donnée ou d'un domaine donné. Mais cette technique ne suffit pas à elle seule quand il s'agit de repérer des collocations équivalentes dans deux langues. Nous avons donc dû avoir recours à des dictionnaires probabilistes.

3.2. Dictionnaire probabiliste

Brown *et al.* (1993) ont développé une série de modèles de traduction probabilistes, dont les paramètres, obtenus automatiquement par entraînement, sont des probabilités conditionnelles P(mc | ms) où *ms* est un mot de la phrase à traduire (mot source) et *mc* est un mot de la phrase traduite (mot cible). Nous avons utilisé ici leur modèle 2.

Le sens général de la valeur P(mc | ms) est la probabilité d'observer un mot *mc* dans la phrase en L2 étant donné la présence du mot *ms* dans la phrase en L1. Par exemple, si L1 est l'anglais et L2, le français, on peut s'attendre que P(chaise | chair) > P(table | chair) ou que P(chaise | chair) > (tabouret | chair). En d'autres mots, la probabilité que *chaise* soit la traduction de *chair* est plus grande que la probabilité que *table* ou *tabouret* soit la traduction de *chair*.

En utilisant les paramètres de ces modèles, on peut construire de façon totalement automatique un dictionnaire probabiliste (DP) de L1 vers L2. Dans ce dictionnaire un peu particulier (voir un exemple d'entrée au tableau 6), tous les mots de L2 peuvent être, à divers degrés, la traduction d'un mot donné *ms* en L1, bien que dans les faits, seul un petit nombre de mots ont une probabilité significative. Cet ensemble de mots en L2 peut être présenté en ordre décroissant pour un mot donné en L1. C'est en utilisant cette méthode de Brown *et al.* que les chercheurs du RALI ont construit deux DP, l'un français → anglais et l'autre anglais → français.

3.3. Coloc

Notre modèle, *coloc*, utilise ces deux DP pour évaluer la probabilité *P(ms' | ms)*, c'est-à-dire la probabilité qu'un mot *ms'* apparaisse dans la phrase en L1 étant donné qu'un autre mot en L1, soit *ms*, est aussi dans cette même phrase.[4]

Si, par exemple, L1 est le français et L2 l'anglais, nous obtenons des DP les probabilités *P(E | F)* et *P(F | E)*, *E* étant un mot anglais et *F* un mot français, par l'équation suivante:

$$P(F' \mid F) = \text{Somme } \textit{(sur tous les mots } E_i \textit{ de l'anglais)}\ P(F' \mid E_i)\ P(E_i \mid F)$$

En fait, *coloc* fait un aller-retour de L1 à L1 en passant par L2. Pour l'exemple précédent, *coloc* part d'un mot français *F* et le cherche dans le DP français-anglais pour obtenir des équivalents possibles de ce mot en anglais (E_i). *Coloc* prend ensuite chacun de ces mots anglais E_i, les cherche à leur tour dans le DP anglais-français pour extraire une série d'équivalents français possibles *F'* pour chaque mot anglais. Pour chaque paire *F-F'*, *coloc* fait le produit des deux probabilités, $P(F' \mid E_i)\ P(E_i \mid F)$, soit la probabilité que *F'* soit la traduction de E_i multiplié par la probabilité que E_i soit la traduction de *F*. Et puisque chaque paire *F-F'* a pu être récupérée plus d'une fois (c'est-à-dire en passant par diverses valeurs de E_i) *coloc* fait finalement la somme de ces probabilités conjointes pour chaque paire *F-F'*. En pratique, on ne fait la somme que sur les *n* premiers mots anglais (*n* étant fixé à 200), l'anglais agissant comme un métalangage capturant certaines relations sémantiques entre *F* et *F'*.

4. Repérage automatique des collocations

Il nous est apparu évident, après avoir examiné les entrées du DP et les sorties de *coloc*, que ces techniques livrent des résultats qui, traités avec l'information mutuelle par exemple, ouvrent la voie au repérage automatique de collocations dans des bitextes. Pour un mot donné en L1, chacune des deux méthodes donne une liste de mots qui, s'ils sont regroupés entre eux deux à deux, pourraient être des collocations (DP donne une liste en L2 et *coloc*, une liste en L1). Il fallait alors décider d'une stratégie qui produirait facilement des résultats utilisables dans le cadre de la lexicographie bilingue.

Nous décrirons ici, étape par étape, la méthodologie suivie pour extraire des listes de collocations possibles pour le mot *erreur/NomC*.

4.1. Repérage des collocations de *erreur*/NomC en L1

Le tableau 1 présente les 100 premiers *collocatifs possibles*, tels que calculés par *coloc* pour le mot à l'étude, soit *erreur/NomC*.

Tableau 1 : Liste L1-1 pour *erreur/NomC*

erreur/NomC	mal/Adve	monumental/AdjQ	donner/Verb
commettre/Verb	faute/NomC	agir/Verb	réparer/Verb
croire/Verb	erroné/AdjQ	trouver/Verb	à tort/Adve
comprendre/Verb	dire/Verb	route/NomC	maintenant/Adve
faire/Verb	compréhension/NomC	rappeler/Verb	entendre/Verb
avoir/Verb	répéter/Verb	façon/NomC	considérer/Verb
penser/Verb	lacune/NomC	bévue/NomC	relever/Verb
tromper/Verb	abuser/Verb	rectifier/Verb	mieux/Adve
être/Verb	glisser/Verb	en fait/Adve	apporter/Verb
tort/NomC	raison/NomC	souvenir/Verb	député/NomC
ne/Adve	aller/Verb	éviter/Verb	très/Adve
pas/Adve	méprendre/Verb	cas/NomC	doute/NomC
faux/AdjQ	pouvoir/Verb	permettre/Verb	malheureusement/Adve
mauvais/AdjQ	contenir/Verb	tout à fait/Adve	réfléchir/Verb
rendre/Verb	devoir/Verb	falloir|faillir/Verb	ministre/NomC
grave/AdjQ	comporter/Verb	là/Adve	sujet/NomC
estimer/Verb	sembler/Verb	suivre|être/Verb	produire/Verb
corriger/Verb	répréhensible/AdjQ	exact/AdjQ	attribuable/AdjQ
à mon avis/Adve	reconnaître/Verb	remédier/Verb	risque/NomC
bien/Adve	bon/AdjQ	déclaration/NomC	juger/Verb
prendre/Verb	vouloir/Verb	leurrer/Verb	tourner/Verb
savoir/Verb	fait/NomC	mémoire/NomC	songer/Verb
avis/NomC	inexact/AdjQ	entente/NomC	correct/AdjQ
mal/NomC	tenir/Verb	clocher/Verb	injuste/AdjQ
chose/NomC	présenter/Verb	convaincre/Verb	vrai/AdjQ
			[...]

Avant de soumettre la liste L1-1 à l'étape suivante, soit le tri par information mutuelle, nous l'avons simplifiée. En effet, il s'avère que de nombreux mots sur cette liste, souvent des verbes et adverbes très généraux comme *être*, *à mon avis*, *ne* et *pas*, sont inintéressants sur le plan collocationnel. De plus, pour des raisons d'ambiguïtés non-résolues lors de la lemmatisation du corpus, certains mots apparaissent sous forme de doublets. Ainsi, afin d'optimiser les résultats à la prochaine étape, les mots très vagues sont éliminés de L1-1 et les doublets,

séparés. Le tableau 2 présente, pour *erreur/NomC*, les mots que nous avons exclus de l'étape suivante et les doublets qu'il s'est avéré nécessaire de séparer.

Tableau 2 : Modifications apportées à L1-1 pour *erreur/NomC*

Mots exclus (ordre alphabétique)				Doublets divisés
à l'occasion	alors	même	ne	falloir|faillir/Verb
à mon avis	aussi	pas	là	première|premier/NomC
à tout	avoir	plus	peut-être	suivre|être/Verb
en fait	bien	tout	sans doute	
là-dedans	encore	très	tout à fait	
malheureusement	être			

À l'étape suivante (tableau 3), nous avons multiplié la liste L1-2 par elle-même (produit cartésien) pour produire des paires de mots :

Tableau 3 : Produit cartésien de L1-2 x L1-2

	mot_1	mot_2	mot_3	...	mot_j
mot_1	(mot_1, mot_1)	(mot_1, mot_2)	(mot_1, mot_3)	...	(mot_1, mot_j)
mot_2	(mot_2, mot_1)	(mot_2, mot_2)	(mot_2, mot_3)	...	(mot_2, mot_j)
...	...	...	...	...	...
mot_i	(mot_i, mot_1)	(mot_i, mot_2)	(mot_i, mot_3)	...	(mot_i, mot_j)

À chacune de ces paires, on associe une valeur basée sur le calcul de l'information mutuelle [les catégories grammaticales ne sont pas indiquées dans ce tableau, mais il est entendu que chaque paire de mots est, en fait, de forme *(mot_i/cat_gram$_i$, mot_j/cat_gram$_j$)*]. Une fois les paires triées en fonction de cette valeur, elles constituent la liste L1-3. Le prochain tableau présente les 100 premières combinaisons calculées pour le mot *erreur/NomC*.

Tableau 4 : Liste L1-3 des 100 premières paires obtenues pour le mot *erreur/NomC* après le calcul de l'information mutuelle

député/NomC	pouvoir/Verb	répréhensible/AdjQ	incorrect/AdjQ
pouvoir/Verb	député/NomC	faire/Verb	savoir/Verb
croire/Verb	député/NomC	commettre/Verb	erreur/NomC
monsieur/NomC	façon/NomC	dire/Verb	donner/Verb
député/NomC	croire/Verb	penser/Verb	chambre/NomC
pouvoir/Verb	compte/NomC	gouvernement/NomC	dernier/AdjQ
tourner/Verb	rond/AdjQ	apporter/Verb	correction/NomC
compte/NomC	pouvoir/Verb	ministre/NomC	question/NomC
cas/NomC	question/NomC	gouvernement/NomC	pouvoir/Verb
question/NomC	cas/NomC	monsieur/NomC	bon/AdjQ
gouvernement/NomC	mesure/NomC	corriger/Verb	lacune/NomC
gouvernement/NomC	mettre/Verb	erreur/NomC	glisser/Verb
bévue/NomC	monumental/AdjQ	ministre/NomC	concerner/Verb
mesure/NomC	gouvernement/NomC	commettre/Verb	monumental/AdjQ
mémoire/NomC	fidèle/AdjQ	devoir/Verb	bon/AdjQ
façon/NomC	monsieur/NomC	réparer/Verb	bévue/NomC
mettre/Verb	gouvernement/NomC	gouvernement/Nom	suivre/Verb

bévue/NomC	rectifier/Verb	fautif/AdjQ	inexact/AdjQ
aller/Verb	tenir/Verb	correction/NomC	incorrect/AdjQ
échec/NomC	monumental/AdjQ	député/NomC	sujet/NomC
commettre/Verb	bévue/NomC	simplement/Adve	gouvernement/NomC
tenir/Verb	aller/Verb	répréhensible/AdjQ	fautif/AdjQ
faire/Verb	comprendre/Verb	député/NomC	dire/Verb
devoir/Verb	faire/Verb	dire/Verb	peut-être/Adve
savoir/Verb	faire/Verb	député/NomC	grand/AdjQ
comprendre/Verb	faire/Verb	éviter/Verb	malentendu/NomC
prendre/Verb	mesure/NomC	glisser/Verb	oubli/NomC
inexact/AdjQ	incorrect/AdjQ	oubli/NomC	glisser/Verb
erreur/NomC	monumental/AdjQ	trouver/Verb	redire/Verb
question/NomC	gouvernement/NomC	ministre/NomC	faire/Verb
ministre/NomC	penser/Verb	rectifier/Verb	erroné/AdjQ
gouvernement/NomC	question/NomC	pouvoir/Verb	faire/Verb
bévue/NomC	passé/AdjQ	question/NomC	mesure/NomC
rectifier/Verb	oubli/NomC	devoir/Verb	rendre/Verb
monsieur/NomC	faire/Verb	inexact/AdjQ	erroné/AdjQ
comporter/Verb	lacune/NomC	pouvoir/Verb	gouvernement/NomC
penser/Verb	ministre/NomC	ministre/NomC	venir/Verb
comporter/Verb	faille/NomC	gouvernement/NomC	simplement/Adve
ministre/NomC	laisser/Verb	monsieur/NomC	pouvoir/Verb
dire/Verb	droit/NomC	faire/Verb	devoir/Verb
question/NomC	ministre/NomC	bévue/NomC	faute/NomC
juger/Verb	correct/AdjQ	question/NomC	député/NomC
réparer/Verb	tort/NomC	dire/Verb	partie/NomC
correct/AdjQ	fautif/AdjQ	corriger/Verb	erreur/NomC
bévue/NomC	commettre/Verb	bon/AdjQ	devoir/Verb
mesure/NomC	prendre/Verb	voir/Verb	gouvernement/NomC
dernier/AdjQ	gouvernement/NomC	erreur/NomC	commettre/Verb
député/NomC	question/NomC	devoir/Verb	pouvoir/Verb
réparer/Verb	oubli/NomC	remédier/Verb	lacune/NomC
ministre/NomC	gouvernement/NomC	reconnaître/Verb	pouvoir/Verb

Certes, cette liste est intéressante telle quelle, mais, puisqu'en lexicographie, on travaille sur un mot donné, en l'occurrence *erreur/NomC*, il sera utile de n'afficher que les paires qui contiennent effectivement *erreur/NomC*. En balayant L1-3 à l'aide de la commande Unix *grep*, il est facile d'en extraire toutes les paires qui contiennent le mot *erreur/NomC*, comme en fait foi le tableau 5.

Tableau 5 : Liste L1-4 des combinaisons en L-1contenant le mot *erreur/NomC*

erreur/NomC	monumental/AdjQ	erreur/NomC	commettre/Verb
commettre/Verb	erreur/NomC	rectifier/Verb	erreur/NomC
erreur/NomC	glisser/Verb	erreur/NomC	passé/AdjQ
corriger/Verb	erreur/NomC	bévue/NomC	erreur/NomC

Si ces résultats semblent contenir des répétitions, c'est que l'information mutuelle se calcule pour un couple ordonné. Ainsi, *commettre + erreur* n'est pas équivalent à *erreur + commettre* (cette dernière représentant sans doute la forme passive). Selon nous, des huit combinaisons repérées, deux seulement, soit *erreur + passé* et *bévue + erreur*, ne sont pas des collocations puisque, dans le premier cas, *erreur* peut être associé à toute la série de synonymes ou de quasi synonymes de l'adjectif *passé* (comme *précédent*, *antérieur* et *historique*) et, dans le second, *bévue* et *erreur* sont des synonymes.

4.2. Repérage de collocations en L2 associées à *erreur/NomC*

Si l'utilisation de *coloc* suivie du calcul de l'IM permet d'extraire des collocations possibles d'*erreur/NomC*, ce qui suit décrit de quelle façon nous avons obtenu une liste d'équivalents potentiels aux collocations en L1 déjà extraites.

Dans un premier temps, il faut extraire du DP la liste L2-1 des traductions possibles d'*erreur/NomC* (tableau 6).

Tableau 6 : Liste L2-1 pour *erreur/NomC*

mistake/NomC	oversight/NomC	stand/Verb	true/Adve
error/NomC	recall/Verb	miscalculation/NomC	context/NomC
make/Verb	miscarriage/NomC	erroneously/Adve	big/AdjQ
wrong/Adve	fault/NomC	drafting/NomC	move/NomC
believe/Verb	erroneous/AdjQ	misunderstanding/NomC	submission/NomC
think/Verb	in fact/Adve	terrible/AdjQ	may/Verb
understand/Verb	case/NomC	part/NomC	point/NomC
be/Verb	clerical/AdjQ	do/Verb	flawed/AdjQ
mistake/Verb	failure/NomC	back/Adve	argue/Verb
understanding/NomC	now/Adve	somewhere/Adve	house/NomC
wrong/AdjQ	correct/AdjQ	wrong/NomC	piece/NomC
correct/Verb	statement/NomC	past/NomC	completely/Adve
not/Adve	lead/Verb	miss/Verb	bank/NomC
err/Verb	memory/NomC	minister/NomC	somehow/Adve
mistaken/AdjQ	fallacy/NomC	accurate/AdjQ	mentality/NomC
right/Adve	inaccurate/AdjQ	track/NomC	innocent/AdjQ
incorrect/AdjQ	remember/Verb	occur/Verb	you're/Verb
flaw/NomC	fact/NomC	simply/Adve	blunder/NomC
way/NomC	gather/Verb	possibility/NomC	thing/NomC
have/Verb	egregious/AdjQ	false/AdjQ	guess/Verb
correctly/Adve	repeat/Verb	unfortunately/Adve	admit/Verb
bad/AdjQ	factual/AdjQ	then/Adve	place/NomC
will/Verb	misconception/NomC	out/Adve	suggest/Verb
say/Verb	prove/Verb	proposition/NomC	indeed/Adve
grievous/AdjQ	opinion/NomC	serve/Verb	inadvertently/Adve
			[...]

Ce qui frappe d'abord dans cette liste L2-1, ce sont les équivalents possibles d'*erreur/NomC*, notamment *mistake/NomC*, *error/NomC*, *flaw/NomC*, *oversight/NomC* et *fault/NomC*. Bien entendu, ces équivalents ne traduisent pas *erreur* dans tous les contextes, mais ils donnent néanmoins une bonne idée de la panoplie de traductions possibles. De plus, certaines combinaisons possibles entre les mots figurant sur cette liste, comme *make* + *mistake*, *make* + *error*, *correct* + *mistake* et *grievous* + *mistake*, sautent aux yeux. Et comme c'était le cas avec les résultats de *coloc* en L1, la liste L2-1 doit être modifiée afin d'optimiser les résultats à l'étape suivante. Puisqu'il ne semble pas y avoir eu de problèmes d'ambiguïté au niveau de la lemmatisation, il suffit d'éliminer les mots trop vagues ou trop généraux de L2-1 avant de passer au produit cartésien de cette liste par elle-même, puis au traitement par IM.

Le tableau 7 présente les mots qui ont été exclus de L2-1 pour optimiser le calcul de l'information mutuelle.

Tableau 7 : Mots exclus de la liste L2-1 pour *erreur/NomC*

back	ndeed	out	then
be	may	simply	unfortunately
have	not	somehow	will
in fact	now	somewhere	you're

La nouvelle liste L2-2, qui ne contient qu'une série de *mot/cat_gram*, sera maintenant soumise au calcul de l'information mutuelle.

Comme c'était le cas en L1, le modèle effectue d'abord le produit cartésien de la liste L2-2 par elle-même pour former une nouvelle liste qui contient des couples de type *(mot$_i$/cat_gram$_i$, mot$_j$/cat_gram$_j$)*. Chaque couple sur cette liste est ensuite associé à une valeur d'information mutuelle, comme le montre la liste L2-3 pour *erreur/NomC* (tableau 8).

Tableau 8 : Liste L2-3 des 100 premières paires obtenues pour *erreur/NomC* après le calcul de l'information mutuelle

goof/NomC	goof/Verb	mistake/NomC	blunder/NomC
egregious/AdjQ	goof/NomC	think/Verb	make/Verb
think/Verb	government/NomC	false/AdjQ	untrue/NomC
egregious/AdjQ	error/NomC	paper/NomC	clerical/AdjQ
fallacy/NomC	mistakenly/Adve	oversight/NomC	drafting/NomC
government/NomC	pass/Verb	drafting/NomC	error/NomC
sin/NomC	omission/NomC	statement/NomC	do/Verb
government/NomC	think/Verb	again/Adve	member/NomC
mistake/Verb	untrue/NomC	sin/NomC	wrong/NomC
grievous/AdjQ	error/NomC	untrue/NomC	mislead/Verb
pass/Verb	government/NomC	err/Verb	side/NomC
inadvertently/Adve	incorrectly/Adve	clerical/AdjQ	oversight/NomC
mistakenly/Adve	execute/Verb	right/Adve	wrong/NomC
quasi-judicial/AdjQ	body/NomC	presume/Verb	innocent/AdjQ
clerical/AdjQ	error/NomC	dead/Adve	wrong/Adve
factual/AdjQ	erroneously/Adve	incorrectly/Adve	exact/AdjQ
sin/NomC	time-consuming/AdjQ	member/NomC	again/Adve
inadvertently/Adve	mislead/Verb	inadvertent/AdjQ	error/NomC
think/Verb	people/NomC	goof/NomC	mistake/NomC
gather/Verb	glitch/NomC	mistake/NomC	miscalculation/NomC
erroneously/Adve	inadvertently/Adve	inadvertent/AdjQ	mislead/Verb
inadvertently/Adve	erroneously/Adve	say/Verb	go/Verb
fallacy/NomC	misconception/NomC	correct/Verb	error/NomC
presume/Verb	goof/Verb	mislead/Verb	inaccurate/AdjQ
drafting/NomC	oversight/NomC	mistake/NomC	goof/Verb
mark/NomC	goof/NomC	flawed/AdjQ	erroneous/AdjQ
innocent/AdjQ	execute/Verb	egregious/AdjQ	mistake/NomC
erroneously/Adve	execute/Verb	correction/NomC	offender/NomC
detect/Verb	wrong/NomC	inadvertent/AdjQ	mistake/NomC
detect/Verb	wrong/NomC	true/Adve	untrue/NomC
wrong/NomC	quasi-judicial/AdjQ	inadvertently/Adve	misunderstand/Verb
error/NomC	omission/NomC	totally/Adve	inaccurate/AdjQ
memory/NomC	correctly/Adve	honest/AdjQ	mistaken/AdjQ
execute/Verb	innocent/AdjQ	accurate/AdjQ	factual/AdjQ
incorrectly/Adve	conceive/Verb	grievous/AdjQ	mistake/NomC
goof/Verb	okay/Adve	totally/Adve	erroneous/AdjQ
okay/Adve	goof/Verb	mistakenly/Adve	accident/NomC
minister/NomC	make/Verb	mistaken/AdjQ	mentality/NomC
big/AdjQ	goof/NomC	error/NomC	oversight/NomC
conceive/Verb	blunder/NomC	erroneous/AdjQ	incorrect/AdjQ

goof/NomC	error/NomC	make/Verb	minister/NomC
misconception/NomC	mistaken/AdjQ	detect/Verb	shortcoming/NomC
innocent/AdjQ	mistakenly/Adve	detect/Verb	shortcoming/NomC
erroneous/AdjQ	error/NomC	wrong/NomC	execute/Verb
do/Verb	statement/NomC	grievous/AdjQ	terrible/AdjQ
people/NomC	think/Verb	terrible/AdjQ	grievous/AdjQ
factual/AdjQ	error/NomC	erratic/AdjQ	occur/Verb
guess/Verb	erratic/AdjQ	presume/Verb	inadvertent/AdjQ
error/NomC	grievous/AdjQ	error/NomC	egregious/AdjQ
correct/Verb	wrong/NomC	dead/Adve	track/NomC

Cette liste montre bien des équivalents potentiels pour les collocations déjà repérées en français, mais elles y sont présentées pêle-mêle. Pour faciliter l'établissement de corrélations entre L1-4 (liste des collocations en L1) et L2-4, les couples constituant L2-4 ont été regroupés autour des substantifs, puisque le mot *erreur*, qui fait l'objet de cette étude, est un substantif (tableau 9).

Tableau 9 : Liste L2-4 pour *erreur/NomC* regroupée autour des substantifs (extrait)

error		**mistake**	
egregious/AdjQ	error/NomC	mistake/Verb	untrue/NomC
grievous/AdjQ	error/NomC	mistake/NomC	blunder/NomC
clerical/AdjQ	error/NomC	goof/NomC	mistake/NomC
error/NomC	omission/NomC	mistake/NomC	miscalculation/NomC
goof/NomC	error/NomC	mistake/NomC	goof/Verb
erroneous/AdjQ	error/NomC	egregious/AdjQ	mistake/NomC
factual/AdjQ	error/NomC	inadvertent/AdjQ	mistake/NomC
error/NomC	grievous/AdjQ	grievous/AdjQ	mistake/NomC
drafting/NomC	error/NomC	mistake/NomC	inadvertent/AdjQ
inadvertent/AdjQ	error/NomC	mistake/Verb	misguided/AdjQ
correct/Verb	error/NomC	**oversight**	
error/NomC	oversight/NomC	drafting/NomC	oversight/NomC
error/NomC	egregious/AdjQ	oversight/NomC	drafting/NomC
misconception/NomC	error/NomC	clerical/AdjQ	oversight/NomC
error/NomC	erroneously/Adve	correct/Verb	oversight/NomC
error/NomC	incorrectly/Adve	**goof**	
error/NomC	blunder/NomC	goof/NomC	goof/Verb
error/NomC	drafting/NomC	egregious/AdjQ	goof/NomC
error/NomC	miscalculation/NomC	mark/NomC	goof/NomC
error/NomC	erroneous/AdjQ	big/AdjQ	goof/NomC
error/NomC	execute/Verb	**blunder**	
flaw		conceive/Verb	blunder/NomC
correct/Verb	flaw/NomC	[...]	

4.3. Mise en correspondance des collocations en L1 et en L2

Maintenant que les listes L1-4 et L2-4 sont triées, il est maintenant plus facile d'établir des équivalences entre les deux listes de collocations (tableau 10).

Tableau 10 : Equivalences possibles entre L1-4 et L2-4 pour le mot *erreur/NomC*

Collocations L1	Collocations L2	Collocations L1	Collocations L2
erreur + monumental	grievous + error error + grievous grievous + mistake egregious + error error + egregious egregious + mistake egregious + goof	corriger + erreur rectifier + erreur	correct + error correct + flaw correct + oversight
commettre + erreur	make + mistake error + execute	erreur + glisser	–

5. Critique

Force est d'admettre que ces listes sont loin d'être exhaustives. En effet, seules cinq collocations françaises ont été repérées (six si l'on inclut *erreur + passé*), et une (*erreur + glissée*) est restée sans équivalent. Qu'est-il advenu de collocations tout à fait habituelles comme *tomber + erreur* et *induire + erreur*? Pourquoi les combinaisons *commit + mistake* ou *mistake + slip* n'ont-elles pas été repérées?

La raison est simple. Les listes L1-1 et L2-1 utilisées pour le produit cartésien comprenaient (dans un cas comme dans l'autre) moins de 200 mots. Ainsi, les verbes *induire*, *tomber*, *slip* et *commit*, ne figurant pas sur ces premières listes, n'ont pas fait partie des calculs subséquents et ne sont donc pas apparus sur les listes de collocations.

Nous tentons présentement de déterminer combien de mots des listes L1-1 et L2-1 devraient, idéalement, être utilisés dans le calcul du produit cartésien pour maximiser le nombre de collocations repérées en L1 et en L2.

Bien entendu, les données extraites dépendent entièrement de la nature même des textes. Puisque le bitexte exploité ici est un corpus de nature plutôt orale[5], les collocations repérées ne sont pas aussi riches et nombreuses que si le bitexte avait été constitué d'un mélange de textes oraux et écrits (journalistiques, littéraires, scientifiques, etc.). De plus, de par la nature particulière de notre bitexte, certaines collocations moins fréquentes risquent de ne pas y apparaître. Idéalement, nos calculs devraient être faits sur un bitexte assez varié dans sa composition pour être un peu plus représentatif de la langue prise dans son ensemble et assez volumineux pour se prêter aux méthodes probabilistes.

Cela dit, si la technique de repérage automatique de collocations bilingues dans des bitextes fonctionne sur le Hansard, elle fonctionnera aussi sur d'autres bitextes, dans la mesure où ceux-ci seront préparés et traités de la même façon que le bitexte utilisé ici.

En plus de proposer des équivalents aux collocations repérées, les techniques que nous avons utilisées donnent un *poids* aux collocations, même si nous avons omis ces valeurs de nos tableaux pour en faciliter la lecture. Cette pondération permet au lexicographe de saisir plus facilement ce qui est central et de ce qui est périphérique, une distinction particulièrement

importante dans le domaine des collocations où il est facile de surestimer ce que l'on connaît et, à l'inverse, de mésestimer ou de passer outre les collocations qu'on ne connaît pas.

6. Notes

[1] Nous voudrions remercier nos lecteurs (Benoît Habert, Roda P. Roberts et les évaluateurs de notre article) pour leurs précieux commentaires.

[2] Voir Smadja *et al.* (1996) pour une description de Champollion.

[3] Le Hansard a été aligné par les chercheurs du RALI (laboratoire de Recherche appliquée en linguistique informatique) de l'Université de Montréal (anciennement le CITI).

[4] Puisque nous cherchons présentement à optimiser les résultats en variant divers paramètres, nous ne décrirons pas ici les calculs effectués pour obtenir les résultats actuels.

[5] Les textes en langue source, qui sont oraux au départ, sont transcrits sans beaucoup de modifications, tandis que les textes en langue d'arrivée sont produits par des traducteurs professionnels.

7. Bibliographie

Benson, M., Benson, E. et R. Ilson (1986). *Dictionary of English: A Guide to Word Combinations*, John Benjamins, Amsterdam/Philadelphie.

Brown, P.F., Della Pietra, S.A., Della Pietra, V.J. et R.L. Mercer (1993). The Mathematics Statistical Machine Translation : Parameter Estimation , in *Computational Linguistics*, Vol. 19,n° 2, pp. 263-311.

Church, K.W. et P. Hanks (1990). Word Association Norms, Mutual Information and Lexicography, *Computational Linguistics*, Vol. 16, n° 1, pp. 22-29.

Cowie, A.P. (1978). The Place of illustrative material and collocations in the design of a learner's dictionary, in *In Honor of A.S. Hornby* (ed. P. Strevens), Oxford University Press, Oxford, pp. 149-165.

Crystal, D. (1991). *A Dictionary of Linguistics and Phonetics*, Blackwell Reference, Oxford/Cambridge, 3e édition.

Hausmann, F.J. (1990). Le dictionnaire de collocations, in *Dictionaries: An International Encyclopaedia of Lexicography* (éds. F.J. Hausmann *et al.*), Walter de Guyter , Berlin/New York, pp. 1010-1019.

Isabelle, P. et S. Warwick-Armstrong (1993). Les corpus bilingues : une nouvelle ressource pour le traducteur. *La Traductique* (eds. P. Bouillon et A. Clas), Les Presses de l'Université de Montréal, Montréal, pp. 288-308.

Liang, S.Q. (1991). À propos du dictionnaire français-chinois des collocations françaises, in *Cahiers de lexicologie*, n° 59, pp. 151-167.

McEnery, T. et A. Wilson (1996). *Corpus Linguistics*, Edinburgh University Press, Édimbourg.

Simard, M., Foster, G. et P. Isabelle (1992). Using Cognates to Align Sentences in a Bilingual Corpus, in *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 67-81.

Sinclair, John (éd) (1987). *Looking Up - An Account of the COBUILD Project in Lexical Computing*, HarperCollins, Londres.

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

Smadja, F., McKeown, K.R. et V. Hatzivassiloglou (1996). Translating Collocations for Bilingual Lexicons: A Statistical Approach, in *Computational Linguistics*, Vol. 22, n° 1, p. 1-38.