Université d'Ottawa · University of Ottawa

Bitexte, bi-concordance et collocation

© par Lucie Langlois

sous la direction de Madame Roda P. Roberts

Thèse présentée à

l'École des études supérieures et de la recherche

de l'Université d'Ottawa

pour l'obtention de la Maîtrise en Traduction

Université d'Ottawa

École de traduction et d'interprétation

À ma mère, Viviane Gignac, et à mon père, feu Jean Paul Langlois,
qui furent mes premiers mentors, et à Mars, qui fut mon dernier...

ABSTRACT

Corpus analysis has become an essential part of unilingual lexicography and its usefulness has been shown time and again. Unfortunately, few researchers have actually written about the usefulness of bilingual corpora in bilingual lexicography, perhaps because bitext and bilingual concordancers are still quite new and quite rare. This thesis will try to bridge the gap.

In writing this thesis, we had two main objectives: to show the importance of bilingual corpora in bilingual lexicography and to design a program that extracts collocations and their possible translations.

We will thus highlight the role and usefulness of corpora -- and of bilingual corpora in particular -- in bilingual lexicography first by clarifying the terminology and then by showing how unilingual and bilingual corpora are used differently.

The second part of the thesis deals with the automatic extraction of collocations from a bitext. We will show how bi-concordancers can be useful in extracting collocations, present two programs specially designed to extract collocations and propose a method to automatically extract collocations from bitext. Finally, we will compare our results to those extracted by software that already exists.

REMERCIEMENTS

Beaucoup de personnes ont participé, de près ou de loin, à l'élaboration de cette thèse. Je ne peux toutes les nommer ici, mais je leur offre mes remerciements les plus sincères.

Cependant, je tiens à souligner plus particulièrement le soutien de quelques proches. Grand merci à mon mari, Robert Bedford, qui, armé d'une patience indéfectible, m'a appuyée et secondée dans le quotidien, sans jamais perdre espoir. C'est lui qui m'a encouragée à retourner aux études puis à entreprendre une maîtrise. Merci à mes fils extraordinaires, Jameson et Alexandre Bevington, qui, même s'ils ont dû partager leur maman avec un ordinateur et une pile de livres pendant quelques années, n'ont jamais cessé de me démontrer leur amour sans bornes (chez nous, par contre, le mot *thèse* est maintenant tabou...!) Je veux aussi mentionner le dévouement de mon amie, Joan Goguen, qui m'a remplacée à la maison pendant mes absences.

Je m'en voudrais de ne pas nommer ici certains membres de l'équipe TAO du CITI. Un gros merci d'abord à Pierre Isabelle, qui a cru en ce projet, à Michel Simard, mon manuel Unix interactif, et à Pierre Plamondon, qui a travaillé avec moi dans l'élaboration de divers programmes.

Toute ma gratitude, enfin, à Clara Foz, qui a toujours été là.

Je tiens aussi à remercier le CRSH et BESO de leur appui financier pendant mes études.

Finalement, je ne peux passer sous silence le travail, la patience, les conseils, l'entrain, l'indulgence et la confiance de ma directrice de thèse, Madame Roda P. Roberts. Sans elle, rien n'aurait été possible.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

INTRODUCTION

0.1 Les ressources du lexicographe

Trois principales sources d'information sont à la disposition du lexicographe : les dictionnaires existants, l'introspection et l'observation de la langue en usage.

En règle générale, un lexicographe rédige une entrée de dictionnaire après avoir consulté d'autres dictionnaires : chaque entrée résultant d'une analyse en profondeur de la part d'un autre lexicographe, il serait malheureux de négliger cette source d'information. Dans cette optique, les dictionnaires consultés doivent être des ouvrages sérieux[1], mis à jour régulièrement.

Lorsqu'il rédige une entrée, un lexicographe doit aussi se fier à son intuition, ce qu'on appelle l'*introspection*. Ainsi, le langagier s'appuie sur ses connaissances formelles et intuitives de la langue lorsqu'il s'interroge sur le sens et l'utilisation d'un mot donné.

Finalement, aucune entrée ne devrait être rédigée sans que le lexicographe ait *observé* la langue en usage. Si, historiquement, cette observation a pu se limiter à ce que des lecteurs, souvent bénévoles, repéraient au fil de leurs lectures, comme c'était le cas du Webster's Third, elle est aujourd'hui, grâce à l'informatique, beaucoup plus systématique, exhaustive et objective. En

---

[1] La méthodologie utilisée se doit d'être scientifique et l'information présentée dans les entrées doit être présentée de façon systématique.

effet, lorsque le lexicographe prépare une entrée, il interroge un corpus à l'aide d'un concordancier, et l'ordinateur affiche une série de lignes de concordance pour le mot demandé. Ces concordances facilitent la tâche du lexicographe. Est-ce que tel ou tel mot est effectivement utilisé? Certaines divisions sémantiques devraient-elles être éliminées ou ajoutées? Comment les différents sens de ce mot devraient-ils être ordonnés? Quelles sont les combinaisons les plus courantes de ce mot? Autant de questions auxquelles les concordances permettent de répondre plus facilement.

L'utilisation de 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 (John Sinclair 1987a) lors de la rédaction du dictionnaire Collins Cobuild. Cependant, l'incidence des corpus bilingues sur la lexicographie bilingue n'a pas encore fait l'objet de recherches poussées, sans doute à cause de la rareté des corpus bilingues et des outils conçus pour les exploiter (Tony McEnery et Andrew Wilson 1996:129).

0.2 Sujet de la thèse

L'analyse de corpus est devenue une étape essentielle de la lexicographie unilingue, et son utilité n'est plus à démontrer. Malheureusement, peu de chercheurs se sont attardés à prouver l'utilité des corpus bilingues en lexicographie bilingue, peut-être parce que les bitextes et les concordanciers bilingues sont encore très rares. La présente thèse vise à combler quelque peu cette lacune.

0.3 Objectifs

Ce travail comporte deux grands objectifs : faire valoir l'importance des corpus bilingues en lexicographie bilingue et formaliser les règles nécessaires à l'élaboration d'un logiciel conçu pour l'extraction de collocations et de leurs traductions possibles.

Ainsi, nous mettrons en lumière le rôle et l'utilité des corpus, et plus particulièrement des corpus bilingues, en lexicographie bilingue en clarifiant la terminologie liée à l'exploitation des corpus et en montrant la différence entre la façon dont les corpus unilingues et les corpus bilingues peuvent être utilisés.

Dans un deuxième temps, nous discuterons de l'extraction automatique des collocations à partir d'un bitexte. Ainsi, nous montrerons comment les bi-concordances peuvent servir au repérage de collocations, nous présenterons ensuite deux logiciels conçus spécialement pour le repérage des collocations et, finalement, nous proposerons une méthode d'extraction automatique de collocations dans un bitexte et la comparerons à celle qui existe déjà.

0.4 Méthodologie de la recherche

Nous avons tenté, en premier lieu, de faire un tour d'horizon de la documentation théorique dans le but, surtout, de définir les notions avec lesquelles nous allions traiter. Il aurait été difficile de réunir une bibiliographie exhaustive de la question, surtout dans le domaine unilingue, puisque

nombreux sont les chercheurs qui se sont attardés à la question des corpus, des concordances et des collocations. Cela dit, nous croyons que les articles cités dans le cadre de ce travail sont représentatifs des recherches actuelles.

Une fois certaines définitions établies, nous avons analysé quelques logiciels, d'abord un bi-concordancier et ensuite deux extracteurs de collocations. Finalement nous avons proposé une méthode originale visant l'extraction automatique de collocations et de leurs traductions possibles.

0.5 Résumé de la thèse

Dans le premier chapitre, nous nous attarderons à la question des corpus et des bitextes. Nous tenterons d'en arriver à une définition de *corpus* qui soit utile et pertinente dans le cadre de nos travaux en lexicographie. Nous décrirons quelques-unes des nombreuses typologies qui ont été utilisées par divers chercheurs, montrerons en quoi la taille d'un corpus est un critère particulièrement significatif en lexicographie et expliquerons pourquoi il s'avère parfois utile de subdiviser un corpus en sous-corpus. Nous définirons ensuite ce qu'on entend généralement par corpus bilingue et par bitexte. Finalement, nous présenterons quelques-uns des algorithmes d'alignement qui ont été utilisés par divers chercheurs pour apparier des corpus de traduction, en particulier celui qui a été exploité pour créer le bitexte du *Hansard* qui a servi à nos travaux.

Au chapitre 2, il sera question de concordance et de bi-concordance. Nous tenterons de définir ces termes, puis étudierons des concordances produites par un concordancier unilingue, soit PAT.

Nous enchaînerons avec la description du bi-concordancier TransSearch, suivie d'une section dans laquelle nous proposerons diverses façons dont les bi-concordances produites par ce logiciel peuvent être utilisées en lexicographie bilingue.

La première partie du 3e chapitre brossera un tableau de la problématique des collocations. Nous donnerons d'abord un aperçu de quelques-unes des diverses terminologies utilisées, passerons ensuite en revue différentes définitions du terme *collocation* puis discuterons de diverses typologies utilisées par certains chercheurs. Nous comparerons la collocation à diverses autres combinaisons de mots, notamment à l'expression figée, à la combinaison libre et au composé. Nous aborderons le rôle de la collocation dans le dictionnaire bilingue et décrirons finalement deux *extracteurs* de collocations, un qui fonctionne sur un corpus unilingue, *Xtract*, puis l'autre, *Champollion*, qui tourne sur des bitextes.

Nous détaillerons, dans le dernier chapitre, la méthode que nous avons élaborée pour extraire d'un bitexte des collocations et leurs traductions possibles. Nous expliquerons d'abord les divers programmes et modèles utilisés. Nous présenterons ensuite deux cas : le mot français *erreur* et le mot anglais *future*. Pour chaque cas, nous analyserons les résultats produits à chaque étape de notre modèle, puis critiquerons les résultats finals. Nous ferons ensuite le point sur l'informatisation du modèle complet, puis comparerons nos résultats à ceux de Champollion.

Chapitre 1 : *LES CORPUS ET LES BITEXTES*

Introduction

La valeur des corpus en linguistique n'est pas à prouver. En effet, les langues étant particulièrement complexes à décrire, les chercheurs ne peuvent se fier à l'introspection seule pour les étudier. C'est pourquoi l'étude des corpus présente de nombreux avantages, que Jan Svartvik (1992:8) et Geoffrey Leech (1992:106) résument bien.

Selon eux, la linguistique des corpus se préoccupe de performance linguistique[1] plutôt que de compétence linguistique[2]. Les corpus sont, en quelque sorte, des traces concrètes de la langue en usage; ainsi, la linguistique des corpus permet aux chercheurs de formuler des énoncés plus objectifs que ceux qu'ils auraient formulés uniquement par introspection ou par consultation de locuteurs natifs. Les faits révélés par les corpus sont vérifiables, critère essentiel de toute étude qui se veut scientifique. De plus, l'étude des corpus favorise la description linguistique plutôt que la recherche d'universaux, ce qui n'exclut pas pour autant la théorisation. En effet, l'observation de textes et la description linguistique sont essentielles à l'élaboration d'une théorie. Par ailleurs, l'étude des corpus permet aux linguistes de retourner à une méthode de travail empirique plutôt que rationaliste. Par empirisme, on entend que l'observation contribue à la théorie, et non

---

[1] Notion qui désigne la mise en oeuvre effective de la compétence linguistique dans les actes de parole (Georges Mounin 1974:253).

[2] Notion fondamentale qui désigne la connaissance implicite qu'un sujet parlant possède sur sa langue (Mounin 1974:75).

l'inverse. Cette démarche empirique est, selon Leech (1992:112), essentielle en linguistique des corpus. Le chercheur observe d'abord l'information trouvée dans les textes, puis il élabore un modèle qui tiendra compte de toute cette information. Ce modèle peut ensuite être vérifié sur d'autres corpus ou encore comparé à des modèles concurrentiels. Ainsi, les résultats empiriques forcent le chercheur à tenir compte de tous les faits de langue. Aucune tricherie possible : tout phénomène observé se doit d'être expliqué. C'est pourquoi un modèle de langue basé sur l'analyse de corpus sera à la fois qualitatif et quantitatif, car le fait de s'intéresser à la quantité, c'est-à-dire de savoir si un phénomène est généralisé ou pas, n'exclut pas d'emblée une étude qualitative. Pour toutes ces raisons, les corpus sont une source essentielle d'information en linguistique, surtout pour certaines branches de la linguistique appliquée, comme la lexicographie, la traduction automatique[3] et la reconnaissance de la parole.

Malgré ces preuves écrasantes en faveur de l'utilisation des corpus, Svartvik (1992:10) nous met en garde contre une dépendance absolue et aveugle à leur égard. En effet, les chercheurs doivent allier recherche empirique et introspection s'ils veulent ébaucher des modèles de langue qui soient vraiment représentatifs et complets. Ils effectuent d'abord une recherche sur les corpus et ils analysent ensuite les résultats avec leurs yeux de langagiers.

---

[3] Selon D.J. Arnold *et al.* (1994) et John Hutchins et Harold Somers (1992), l'utilisation de corpus a modifié l'approche – IM des chercheurs en traduction automatique (TA) et, par extension, en reconnaissance de la parole. Les système de TA plus anciens, comme Systran, se fondaient sur des grammaires programmées dans le système règle par règle. Ensuite, les systèmes de TA, comme TranslationManager, ont utilisé des exemples déjà traduits, et donc tirés de corpus, pour traduire *automatiquement* des textes. Enfin, la dernière génération de systèmes de TA semblent s'appuyer de plus en plus sur des statistiques calculées à partir de corpus énormes.

Si les méthodes des recherches en linguistique ont beaucoup évolué avec l'avènement des corpus informatisés, celles des recherches lexicographiques en ont profité plus encore. En effet, l'exploitation des corpus électroniques est devenue si importante en lexicographie contemporaine qu'elle possède désormais sa propre dénomination en anglais : *corpus lexicography*. Bien que l'intuition du lexicographe et l'utilisation des dictionnaires existants aient encore un rôle à jouer dans la préparation de nouveaux dictionnaires, elles servent plutôt de complément à l'exploitation des corpus. Étant donné que les corpus témoignent de la langue actualisée, le lexicographe, lui, se doit d'expliquer les réalités linguistiques telles qu'elles apparaissent dans les corpus. Selon Michael Rundell et Penny Stock (1992b:22), le lexicographe fait face à « the inescapability of the information it (corpus evidence) presents ». Que le corpus lui révèle des sens jusqu'alors ignorés[4], de nouveaux usages ou des néologismes, le lexicographe doit en tenir compte. En outre, lorsque le corpus ne lui révèle pas explicitement une unité lexicale ou un sens particulier, le lexicographe peut envisager la possibilité de l'omettre, même s'il apparaît dans d'autres dictionnaires[5]. Le lexicographe trouve aussi dans son corpus des exemples « prêts-à-utiliser » ou, à tout le moins, des contextes qui ne nécessitent que de petites modifications pour pouvoir être utilisés comme exemples, ce qui facilite grandement l'étape de la rédaction des entrées. Il est clair, cependant,

---

[4] Lors de l'examen de la concordance sur *gritty*, par exemple, nous avons remarqué deux sens qui n'apparaissaient pas dans les dictionnaires, soit 1) *grim, difficult* avec des exemples comme *gritty realism* et *the gritty texture of everyday life* et 2) *raw, rough around the edges, hard driving, rhythmic, chunky* inféré d'exemples comme *gritty live shows* et *gritty guitar work*.

[5] Dans le passé, les lexicographes se servaient de la nomenclature des autres dictionnaires pour établir la leur, sans nécessairement pouvoir vérifier si tous ces mots étaient bel et bien utilisés. Par exemple, les mots *frozenness* et *frozenly* apparaissent dans quelques dictionnaires unilingues anglais, mais sont absents des corpus du projet de Dictionnaire canadien bilingue (DCB) et ne feront sans doute pas partie de sa nomenclature.

que les corpus ne livrent pas que des renseignements intéressants; il faut que le lexicographe analyse ces données empiriques pour en tirer le meilleur parti possible.

1.1 Corpus et sous-corpus

1.1.1 Corpus

À l'heure actuelle, quoique beaucoup de chercheurs se servent de corpus, la définition de ce terme est loin de faire l'unanimité. En effet, utilisé par les linguistes au cours des quarante dernières années, le terme *corpus* demeure encore et toujours ambigu.

Leech (1991:8) rapporte que, pour les linguistes des années 50, un corpus est « a sufficiently large body of naturally occurring data of the language to be investigated ». Vingt ans plus tard, Mounin (1974:89), par exemple, précisera cette définition de base en indiquant que ces données linguistiques peuvent être des « énoncés écrits ou enregistrés ». Un corpus se limitait donc, jusqu'à l'entrée en jeu de l'ordinateur, à un ensemble de textes, de livres, d'extraits de livres ou d'enregistrements.

Pour les lexicographes de la même époque, un *corpus* se composait de fiches remplies à la main par des lecteurs, souvent bénévoles, qui notaient, au fil de leurs lectures, les mots jugés « intéressants », les contextes d'utilisation et les références. Les mots et les citations ainsi repérés reflétaient presque toujours des usages inhabituels de la langue, puisque l'usage habituel n'attirait

pas l'attention des lecteurs[6]. Et, bien sûr, les choix des lecteurs ne faisaient pas toujours l'unanimité.

Toutefois, les progrès en informatique ont révolutionné la façon dont les corpus sont composés. De vastes quantités de textes sont maintenant disponibles en format électronique. Et si un texte donné n'existe qu'en copie papier, il peut facilement être informatisé à l'aide d'un lecteur optique. Donc, un corpus, à l'époque moderne, se présente généralement sur support électronique. Leech et Steven Fligelstone (1992:115) font ressortir la notion contemporaine d'informatisation des textes et des enregistrements en parlant de *computer corpus*, qu'ils définissent d'ailleurs comme étant des « bodies of natural language material (whole texts, samples from texts, or sometimes just unconnected sentences), which are stored in machine-readable form ». Cependant, Jan Aarts (1990:18), qui définit un corpus comme une collection de textes suivis, ajoute qu'il ne vaut même pas la peine de préciser que ces textes devraient être *machine readable* car, de nos jours, cela va de soi.

Avec l'informatisation, la définition de *corpus* a aussi évolué en lexicographie. Les fiches dont il était jadis constitué sont maintenant remplacées par des textes complets. Selon la lexicographe Antoinette Renouf (1987:1) « a corpus [is] a collection of texts of the written or spoken word, which is stored and processed on computer for purposes of linguistic research ».

---

[6] Par exemple, Murray, du *Oxford English Dictionary*, déplorait le fait qu'il y ait 50 citations pour le mot *abusion* et seulement 5 instances du mot *abuse* dans les citations relevées par ses lecteurs (Rundell et Stock 1992a:13).

Il existe de nombreuses définitions du mot *corpus*. Néanmoins, malgré l'importance que l'on accorde à la notion de corpus comme telle, son rôle demeure vague. Par exemple, rares sont ceux qui mentionnent la finalité propre d'un corpus, comme si un corpus constituait une fin en soi. C'est David Crystal (1991) qui semble en donner la meilleure idée, même s'il ne précise pas que le corpus est, de nos jours, habituellement informatisé. Pour lui, un corpus est « a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about language ». Donc, selon ce linguiste, on peut partir des évidences présentées dans un corpus pour en arriver à une description de la langue ou, inversement, utiliser le corpus pour prouver des hypothèses préalables sur la langue.

Il faut noter aussi que les définitions présentées ci-dessus ne font pas état des types de textes dont sont composés les corpus. On parle vaguement d'*énoncés* (Mounin 1974:89), de *bodies of natural language material* (Leech et Fligelstone 1992:115), de *collection of texts* (Nelson Francis 1992:17), et, encore plus vaguement de *helluva lot of text, stored on a computer* (Leech 1992:106), sans précision aucune quant à leur nature. C'est sans doute parce que le choix des textes dont un corpus est constitué est directement lié à la raison d'être du corpus, qu'il soit utilisé pour la description linguistique, pour la vérification d'hypothèses ou pour la lexicographie. En terminologie, par exemple, Éva Dauphin (1994:17) entend par corpus « un ensemble de textes

homogènes, c'est-à-dire traitant du même domaine, rédigés et utilisés par le même type de personnes et dans des conditions semblables »[7].

En effet, les corpus, qui peuvent contenir des textes de nature très diverses, sont souvent classés en types (ou sous-catégories) de corpus selon la nature des textes. Le tableau 1 ne présente que quelques-unes des classifications proposées par les chercheurs[8], mais il suffira pour illustrer ce point.

---

[7] On remarque qu'il n'est nullement mention, ici, de textes informatisés, puisque les corpus de textes spécialisés sur support informatique sont particulièrement rares, surtout dans des langues autres que l'anglais, et que les terminologues dépouillent encore beaucoup d'ouvrages manuellement.

[8] Pour chaque couple, les deux types de corpus sont présentés en ordre chronologique, soit le type plus ancien avant le plus récent.

Type	Description
Général	textes très variés; contenu général
vs	
Spécialisé	textes choisis en fonction d'un but très précis (étude d'un domaine donné, d'un dialecte, etc.)
Limité	ensemble de textes auquel on n'ajoute jamais de nouveaux textes
vs	
Illimité	ensemble de textes dynamique auquel des textes sont ajoutés au fil des occasions et des besoins
Principal	ensemble de textes créé pour répondre à une série de critères pré-établis; stable, équilibré et immuable; aussi appelé *corpus fermé*
vs	
Supplémentaire	complément du corpus principal servant aux recherches ponctuelles; dynamique et hétérogène; textes changent selon l'évolution de la langue; aussi appelé *corpus variable*
Échantillon[9]	ensemble de textes limité, statique; répond à des exigences pré-établies très précises; plutôt petit; n'est utile que dans son ensemble
vs	
Étalon	ensemble de textes très élaboré[10]; appelé à changer souvent; représente la langue en évolution

Tableau 1 -- *Types de corpus*

Beaucoup de chercheurs ont utilisé d'autres paramètres pour classer les corpus; ces typologies sont toutefois trop nombreuses pour que nous les présentions toutes dans le cadre du présent travail. Pour illustrer cette panoplie, mentionnons à titre d'exemple les classifications relevées par

---

[9] Les termes *corpus étalon* et *corpus d'échantillons* sont mes traductions respectives des termes *monitor corpus* et *sample corpus* proposés par Sinclair (1982:4).

[10] Semblable à l'archive ou à la collection qui seront définis plus loin.

Douglas Biber (1994:179) : *corpus annoté* et *non annoté*; *textes complets* et *extraits de textes*; *textes obtenus de façon aléatoire* et *textes choisis avec soin*; *textes choisis au hasard dans un domaine particulier* et *textes choisis au hasard mais en proportions prédéfinies*.

Pour éviter la confusion créée par toutes ces catégories de *corpus*, certains suggèrent, comme l'expliquent Mitchell Marcus *et al.* (1994:273), que le terme *corpus* soit réservé aux groupes de textes qui ont été choisis de façon très précise pour répondre à des besoins et critères particuliers. De plus, ils proposent d'utiliser les termes *archive* ou *collection*, qui sont des termes beaucoup plus généraux, pour signifier l'accumulation de textes sur ordinateur, au fil des occasions. C'est d'ailleurs la position qu'a adoptée Gregory James dans sa communication lors d'Interlex 9[11]. Quoique ces distinctions soient utiles, les termes *archive* et *collection* ne sont pas encore implantés dans l'usage.

Pour cette thèse, notre définition de *corpus* sera un amalgame des définitions de Renouf et de Crystal. Ainsi, nous entendrons par ce terme *une compilation de textes informatisés, parlés ou écrits, qui servent de point de départ pour la description linguistique ou pour prouver une hypothèse sur la langue*. Cette définition semble assez complète, car elle décrit d'abord de quoi il est question et en précise ensuite le rôle.

---

[11] Séminaire sur la lexicographie tenu à l'université d'Exeter en avril 1995.

Non seulement un corpus doit être représentatif et équilibré, comme le rappelle Rundell (1995:30), mais il doit aussi être choisi en fonction des buts que se sont fixés ses concepteurs. Par exemple, pour vérifier si la forme passive est bel et bien utilisée en français technique et scientifique, il faudra que le corpus contienne, effectivement, des textes techniques et scientifiques français. En outre, puisque la lexicographie générale se doit de refléter la langue telle qu'elle est utilisée dans toutes les situations de communication, il faut que le corpus lexicographique prenne en compte non seulement la langue écrite, mais aussi la langue orale.

La taille du corpus est aussi fonction du but dans lequel il a été établi. Selon la lexicographe Renouf (1987:1), « the larger the amount of data available, the more reliable would be the statements which could be made about language ». Ainsi, un corpus utilisé à des fins lexicographiques doit contenir des dizaines, voire des centaines, de millions de mots si l'on veut établir des statistiques valables sur l'utilisation de ces mots. Selon la loi de Zipf, en effet, les mots les plus courts sont les plus fréquents, et la grande majorité des mots ne sont utilisés que très rarement (Sinclair 1992:390). Le corpus de Brown (Francis 1967:133), par exemple, qui contient un peu plus d'un million de mots, n'est pas adapté à la lexicographie. En effet, 45 p. 100 des mots contenus dans ce corpus n'apparaissent qu'une seule fois[12]. Le tableau 2 illustre bien le rôle que joue la taille du corpus à ce chapitre.

---

[12] Selon Sinclair (1991:35), ces hapax représenteraient plutôt 50 p. 100 de tout corpus.

Mot anglais	Nombre d'occurrences		
	Corpus de Brown[13]	The Gazette[14]	English Canadian Press[15]
Taille approx. (en millions de mots)	**1**	**7**	**130**
blouses	1	19	345
alarming	1	45	782
administration	1	643	9668
horrid	1	7	120
lexicostatic	1	0	0
enterotoxemia	1	0	0
waitress	2	23	635
colloquial	2	3	53

Tableau 2 — *Rôle de la taille des corpus*

Fait intéressant, les mots *blouses*, *alarming*, *administration*, *horrid*, *lexicostatic* et *enterotoxemia* n'apparaissent qu'une seule fois dans le corpus de l'université Brown, et les mots *waitress* et *colloquial* ne reviennent que deux fois chacun. Ainsi, si l'on se fiait uniquement à ces observations, le corpus de Brown pourrait donner l'illusion que des mots plus rares ou spécialisés (comme *lexicostatic*, *enterotoxemia* et *colloquial*) sont utilisés presque aussi fréquemment que des mots de tous les jours (comme *blouses*, *alarming*, *administration* et *waitress*). Néanmoins, une

---

[13] Pour le corpus de Brown, les données sur les mots *blouses*, *waitress* et *colloquial* sont citées par Francis (1967:135) et les données pour *alarming*, *administration*, *horrid*, *lexicostatic* et *enterotoxemia* sont tirées de Rundell et Stock (1992a:12).

[14] Corpus journalistique d'environ 6,7 millions de mots.

[15] Corpus journalistique d'environ 129 millions de mots.

recherche dans un corpus plus volumineux montre qu'il se creuse un écart entre justement les mots courants et les mots spécialisés ou plus rares. Ainsi, *blouses*, *alarming*, *administration* et *waitress* ont été repérés 19, 45, 643 et 23 fois respectivement dans un corpus d'environ 7 millions de mots et 345, 782, 9 668 et 635 fois chacun dans un corpus d'environ 130 millions de mots, tandis que des mots plus rares ou spécialisés comme *lexicostatic*, *enterotoxemia* et *colloquial* sont soit absents soit rarement utilisés dans les grands corpus journalistiques. Les grands corpus révèlent donc ce que le corpus plus modeste cachait.

La taille du corpus est donc particulièrement importante pour les recherches lexicographiques afin d'éviter, justement, que les mots rares ne soient perçus, à cause de la composition du corpus, comme étant plus courants qu'ils ne le sont vraiment. Selon Kenneth Church et Patrick Hanks (1990:26), un corpus d'un million de mots (comme celui de Brown) ne pourrait servir qu'à repérer certains usages de quelques formes fléchies pour environ 4 000 entrées de dictionnaire.

1.1.2 Sous-corpus

On peut traiter les textes dont un corpus est composé de deux façons. D'abord, ils peuvent être pris dans leur ensemble, en bloc, ce qui permet d'en arriver à une vision plus générale de la langue. Ou encore, on peut les diviser en sous-blocs, c'est-à-dire en sous-corpus, selon le type de recherche à laquelle on se livre. Si, par exemple, on veut faire des études comparatives entre divers genres littéraires ou divers types de textes (vérifier l'aire d'utilisation d'un mot, comparer les variantes stylistiques entre divers domaines, etc.), il s'avère utile de sectionner un corpus en

sous-corpus. Ces sous-corpus peuvent être composés selon divers critères, notamment la langue des textes, l'aire géographique d'origine, leurs genres et leurs registres.

Rundell et Stock (1992b:24) illustrent bien l'importance des sous-corpus lorsqu'ils rapportent les résultats de leur étude du mot *different* dans un sous-corpus d'anglais britannique versus un sous-corpus d'anglais américain. L'intuition des lexicographes suggérait que les Britanniques préféraient utiliser la formulation *different from*, et les Américains, *different than*. Toutefois, une recherche dans les sous-corpus britanniques et américains a révélé que les Anglo-saxons, quelle que soit leur origine, utilisent, dans la grande majorité des cas, la forme *different from* plutôt que *different than*.

Un autre exemple particulièrement parlant est celui du mot *manhood*[16] en anglais. En effet, une analyse comparative de deux sous-corpus littéraires a montré que ce mot était utilisé de manière bien différente selon la nature des textes. En littérature générale, le mot *manhood* ne revêt aucun sens spécial; il veut toujours dire *âge d'homme* ou *masculinité*. Toutefois, dans les romans d'amour, le mot *manhood* remplace le mot *pénis* qui choquerait sans doute certaines lectrices. Si ces sous-corpus littéraires avaient été combinés, cet usage particulier du mot *manhood*, noyé dans la masse, aurait pu passer inaperçu.

---

[16] Cet exemple a été présenté par Michael Rundell lors d'Interlex 1995.

Le regroupement de textes en sous-corpus est particulièrement utile en lexicographie générale, car les sous-corpus permettent au lexicographe de distinguer plus facilement les unités lexicales particulières à une aire géographique, à un niveau de langue ou à un domaine de spécialisation donné. Ainsi, au Dictionnaire canadien bilingue, les textes unilingues sont regroupés en sous-corpus selon le type de textes (journalistiques, littéraires, scientifiques) et l'aire géographique (français canadien, français hexagonal, anglais canadien, anglais américain) dans une grande base de données textuelles désignée TEXTUM.

1.2 Corpus bilingues et bitextes

1.2.1 Corpus bilingues

Dès qu'on dispose de deux corpus qui se ressemblent et qui sont dans deux langues différentes, comme c'est le cas dans TEXTUM, on a déjà ce qu'on appelle un corpus bilingue. En effet, les corpus bilingues, que Reinhard Hartmann (1980) propose d'appeler *corpus parallèles*[17], peuvent prendre différentes formes. Hartmann, par exemple, les divise en deux groupes : les corpus constitués de textes de départ et de leurs traductions, ou *corpus de traduction*, et ceux qui comprennent des textes en deux ou plusieurs langues qui, sans être des traductions mutuelles,

[17] C'est aussi la terminologie adoptée par Roda P. Roberts (1996).

fonctionnent de façon semblable au plan de la situation de communication, ou *corpus comparables*[18].

En lexicographie bilingue, les corpus de traduction présentent certains avantages par rapport aux corpus comparables : les textes et leurs traductions sont intimement liés sémantiquement et, donc, chaque notion exprimée dans le texte de départ devrait se retrouver, en théorie, dans le texte d'arrivée. Les corpus comparables, eux, ne sont pas liés au même titre que les corpus de traduction; ainsi, le lexicographe ne peut pas toujours y trouver des équivalents. Néanmoins, ils permettent aux chercheurs d'étudier des textes qui traitent de sujets semblables et qui fonctionnent à peu près de la même manière sur le plan de la communication, sans que l'un ou l'autre soit influencé par la langue de départ, comme cela peut être le cas des traductions proprement dites.

Même si les corpus de traduction sont d'une aide précieuse en lexicographie bilingue, ils sont plutôt rares. En effet, si les auteurs se montrent relativement disposés à mettre leurs textes originaux à la disposition de la communauté scientifique, les cabinets de traduction, eux, peut-être par peur de faciliter la tâche à la concurrence, gardent jalousement leurs traductions. Ainsi, il est particulièrement difficile d'obtenir des textes traduits sur une grande échelle et d'établir un corpus de traduction. C'est néanmoins ce genre de corpus bilingue qui nous intéresse ici.

---

[18] Le terme *comparable corpora* a été utilisé par Alessandro Enea, dans un texte disponible sur le W3 (http://www.ilc.pi.cnr.it/EAGLES/typology/node21.html) pour dénommer les corpus composés de textes rédigés dans différentes langues mais qui se ressemblent quant au contenu et au rôle que ces textes jouent. L'auteur utilise aussi ce terme pour qualifier des textes écrits dans une même langue, mais de variétés différentes (par exemple, le *International Corpus of English*, ou *ICE*, qui contiendra environ un million de mots de chaque variété d'anglais à travers le monde).

1.2.2 Bitexte

Lorsque Brian Harris (1988) a proposé le terme anglais *bitext* -- BH pour désigner le corpus de traduction, il imaginait un corpus dans lequel les segments équivalents d'un texte de départ et de sa traduction seraient explicitement mis en correspondance. Un tel bitexte faciliterait la comparaison entre textes de départ et textes d'arrivée et permettrait de repérer facilement un terme ou un syntagme accompagné de l'unité correspondante dans l'autre langue. Dans l'esprit de Harris, un *bitexte* ne serait pas deux textes comme tels, mais plutôt un seul texte en deux dimensions, chaque dimension étant une langue différente. Pour lui, un bitexte est un « bilingual hypertext stored in such a way that each retrievable segment consists of a segment in one language linked to a segment in the other language which has the same meaning » (Harris 1988).

Si la mise en correspondance, ou *appariement*, manuelle de ces textes est possible (mais combien ardue), l'ordinateur s'acquitte assez facilement de cette tâche. Afin d'illustrer comment des couples *appariés* pourraient se présenter dans le concret, le tableau 3 montre un paragraphe anglais apparié au passage équivalent français (cet exemple est tiré de Michel Simard *et al.* 1992:68).

couple 1	E1 The crisis our farmers are in right now will affect all of us at a certain point in time.	F1 La crise que vivent en ce moment nos agriculteurs se répercutera sur tous et chacun de nous à un certain moment.
couple 2	E2 We are all consumers and we all need a strong and healthy agricultural sector.	F2 Nous sommes des consommateurs. F3 Nous avons tous besoin d'une agriculture saine et forte.
couple 3	E3 I am glad that the Hon. Member for Algoma (Mr. Foster) mentioned figures in his remarks. E4 Otherwise, the Government might have eluded the problem once again.	F4 Heureusement que le député d'Algoma (M. Foster) a mentionné des chiffres dans ses remarques, sans cela ce gouvernement s'en serait sorti en douce encore une fois.
couple 4	E5 The Hon. Member for Algoma suggested Tuesday night that the Government had to take a clear position and make a commitment to assist our farmers before it is too late.	F5 Le député d'Algoma suggérait mardi soir qu'il fallait que le gouvernement se prononce clairement et s'engage à aider nos agriculteurs avant qu'il ne soit trop tard.

Tableau 3 – *Exemple d'appariement de textes (Hansard)*

Bien que ce passage comprenne cinq phrases dans chaque langue, nous sommes loin de la correspondance biunivoque (correspondance de phrase à phrase). En effet, *E1* est bien l'équivalent de *F1*, tout comme *E5* et *F5* d'ailleurs. Néanmoins, la phrase *E2* est traduite par deux phrases françaises *F2* et *F3*, et *E3* et *E4* sont rendues par une seule, soit *F4*. D'où la complexité de l'appariement automatique.

Bien entendu, il ne suffit pas d'avoir une masse de textes traduits pour arriver à repérer des segments dans une langue et leurs équivalents possibles dans l'autre. Encore faut-il que ces textes

soient alignés (ou *appariés*); il faut que les textes en langue de départ et en langue d'arrivée soient transformés en bitexte. Pour y arriver, différentes méthodes existent.

1.2.3 Alignement

Church et William Gale (1991a) ont utilisé la phrase comme unité de segmentation des textes, le paragraphe étant trop vaste, et l'unité sémantique étant trop difficile à définir et à repérer. Le tableau 4 résume les correspondances phrastiques admises par le modèle de Church et Gale (une phrase peut être traduite par une phrase, deux phrases traduites par une, une par deux, deux phrases par deux phrases, une phrase par zéro et zéro par une).

Nombre de phrases		
en L1		**en L2**
1	-->	1
2	-->	1
1	-->	2
2	-->	2
1	-->	0
0	-->	1

Tableau 4 – *Correspondances phrastiques (modèle de Church et Gale)*

Leur méthode utilise la longueur des phrases en nombre de caractères comme seul et unique critère de mise en correspondance puisque, selon eux, il existe un rapport de proportionnalité entre la longueur d'une phrase en langue de départ et la longueur de sa traduction[19].

Dans cette optique, Church et Gale ont d'abord apparié les paragraphes, c'est-à-dire qu'ils ont mis en parallèle les paragraphes équivalents des textes en L1 et ceux en L2. Ensuite, pour chaque paire de paragraphes, le modèle de Church et Gale établit toutes les combinatoires possibles entre les phrases dont ces paragraphes sont composés, dans la mesure où ces combinaisons respectent les correspondances phrastiques admises. Ensuite, il choisit, après avoir effectué certains calculs, laquelle de ces combinaisons produit le meilleur alignement pour le paragraphe donné. Cette méthode semble fonctionner assez bien; les auteurs rapportent un taux de succès de 95,8 p. 100. Cependant, lorsque les combinaisons de phrases sont plus complexes, ce programme, selon Simard *et al.* (1992) n'est pas très robuste : il dérape facilement et, après un dérapage, il n'arrive pas à se resynchroniser.

Les chercheurs du Centre d'innovation en technologies de l'information (CITI) ont conçu, de leur côté, un algorithme d'appariement basé sur les mots apparentés, *cognates* en anglais, pour apparier un bitexte (Simard *et al.* 1992, Pierre Isabelle et Susan Warwick-Armstrong 1993). Par mots apparentés, on entend généralement une paire de mots de langues différentes qui ont une orthographe semblable et qui partagent certains traits sémantiques, p. ex., *list/liste*, *error/erreur*

---

[19] D'autres chercheurs comme Brown *et al.* (1993) partagent cet avis.

et *tax/taxe*. Les chercheurs du CITI ont élargi cette notion de façon à inclure les noms propres (*Ottawa/Ottawa*), les expressions numériques (*$1,250,000/1 250 000 $*) et même la ponctuation (une question en anglais sera vraisemblablement traduite par une question en français et vice-versa).

Le programme fonctionne de la façon suivante. Pour chaque paragraphe, il dresse d'abord deux listes, une constituée des mots anglais du texte, et l'autre des mots français. Ensuite il tente d'établir des correspondances entre la graphie des mots contenus dans chaque liste. Puisqu'il est plus probable que deux phrases qui sont des traductions mutuelles contiennent des formes semblables que des phrases choisies au hasard, cette méthode semble valable à priori. Néanmoins, il n'est pas exclu non plus que deux phrases qui sont d'excellentes traductions l'une de l'autre ne contiennent aucune forme semblable (*c'est l'été / it's summer*). Ou encore que deux phrases, qui sont tout à fait différentes sur le plan sémantique, contiennent des formes semblables qui n'ont aucun trait sémantique en commun (*il est physicien / he's a physician*).

Bien que cette méthode soit très coûteuse en calcul (elle prend neuf fois plus de temps que la méthode basée sur la longueur des phrases), les résultats sont quand même intéressants. En effet, si l'on compare un texte apparié en fonction des formes semblables à un alignement manuel réalisé par huit chercheurs au CITI, elle affiche un taux d'erreur de seulement 2,4 p. 100.

Afin d'être en mesure de mieux évaluer leur approche, les chercheurs du CITI l'ont comparée à celle qui est utilisée par Church et Gale. Ainsi, ils ont appliqué l'algorithme de ces derniers au

même corpus que celui qu'ils avaient eux-mêmes utilisé pour leur propre algorithme et ont obtenu un taux d'erreur de seulement 1,8 p. 100[20].

Aussi le CITI a-t-il décidé de jumeler ces deux méthodes et de se servir des mots apparentés pour améliorer le rendement du modèle de Church et Gale. Ainsi, le nouveau programme fonctionne en deux temps. D'abord, il tente d'aligner les textes en fonction de la longueur des phrases selon le modèle de Church et Gale. Lorsque le programme achoppe sur un passage précis, il retient les deux segmentations les plus probables basées strictement sur le plan de la longueur des segments plutôt que de forcer un choix comme c'était le cas dans l'algorithme original de Church et Gale. Ensuite, le programme utilise les mots apparentés pour préférer une segmentation à l'autre. Cette façon de procéder améliore l'appariement, puisque le taux d'erreur est réduit à 1,6 p. 100.

1.2.4 Hansard

Le bitexte utilisé dans le cadre de cette thèse est le Journal des débats du Parlement canadien, communément appelé le *Hansard*[21]. Ce bitexte, sans doute le plus grand bitexte français/anglais disponible en format électronique aujourd'hui, a été aligné par le CITI selon l'algorithme à deux temps décrit ci-dessus. Si quelques fois le français est la langue de départ, la majorité des textes sont d'abord rédigés en anglais.

---

[20] Ces résultats sont meilleurs que ceux rapportés par Church et Gale.

[21] Il est composé de huit ans de débats (1988 à 1995).

Le Hansard est aussi un texte étiqueté[22]. Ainsi, il permet des recherches non seulement sur les formes lexicales, mais aussi leurs catégories grammaticales. Cette possibilité nous a permis de concentrer nos recherches sur les verbes, les substantifs, les adjectifs et les adverbes, puisque ce sont des mots de ces catégories qui constituent le plus souvent la base des collocations que nous voulions extraire automatiquement.

---

[22] On entend par *corpus étiqueté* un corpus dans lequel chaque mot est accompagné de sa catégorie grammaticale.

Chapitre 2 : *LES CONCORDANCES ET LES BI-CONCORDANCES*

Introduction

Il va sans dire que les corpus unilingues et bilingues ne peuvent être utiles que si l'on peut les exploiter de façon à en extraire l'information voulue. Le présent chapitre traitera de la manière plus *traditionnelle* d'exploiter des corpus, soit par l'utilisation de *concordanciers* pour obtenir des *concordances*.

Comme l'explique Leech (1992:106), un corpus informatisé est, comme tel, de peu d'intérêt. Ce n'est que lorsqu'il est jumelé à un concordancier puissant, c'est-à-dire à un logiciel qui permet de chercher, de repérer et de classer des *mots* ainsi que de calculer et d'afficher des résultats, qu'il peut être exploité à fond. Pour pouvoir tirer profit d'un corpus au maximum, le lexicographe doit être en mesure d'organiser des extraits de corpus selon certains paramètres qui lui permettront, notamment, de distinguer les divers sens des mots, de repérer leurs combinatoires typiques ou encore de voir de quelle façon ces mots fonctionnent sur le plan syntaxique (Simon Baugh *et al.* 1996:40). En d'autre termes, le lexicographe doit pouvoir extraire des concordances qui correspondent à ses besoins précis en utilisant un concordancier interactif.

2.1 Concordance et concordancier

Tout comme le terme *corpus*, la définition de *concordance* a subi quelques modifications au fil du temps. En effet, historiquement, une *concordance* signifiait « an alphabetically ordered list of [...] all words of the text T, where for each token w of T there is given some information about the position of w in T (e.g. book, chapter, verse, page, line) and an appropriate context in which w appears in T » (Wolfgang Slaby 1978:117). Cette liste contenait habituellement tous les contextes de *w* dans T. La concordance d'une oeuvre complète se présentait donc sous la forme d'un livre contenant la liste exhaustive des contextes d'utilisation de chaque mot lexical.

Puisque ces listes étaient originairement produites à la main, elles n'étaient dressées que pour des oeuvres d'envergure, notamment la Bible et les pièces de Shakespeare. En fait, comme le fait remarquer Todd Bender (1977:54), la concordance, à l'époque pré-informatique, était vue comme une fin en soi à cause de l'envergure du projet, plutôt que comme un outil faisant partie de recherches plus ambitieuses. Avec l'avènement de l'ordinateur, toutefois, on a établi des concordances pour un plus grand nombre d'oeuvres et d'auteurs, grâce auxquelles les chercheurs peuvent facilement analyser le style de divers auteurs et comparer les images, les expressions et même les types de verbes que chacun privilégie.

Par ailleurs, de nos jours, la concordance n'est plus forcément rattachée à une oeuvre donnée ou à un auteur particulier. La technologie actuelle permet de chercher des concordances au cas par cas, un mot à la fois, au fil des besoins, sans avoir à consulter la concordance de textes entiers.

Selon Leech (1992:114), une concordance est devenue « a file listing of all (or perhaps a sample) occurrences of the target item, together with sufficient context for the research purpose ». Pour Sinclair (1991:32), une concordance « is a collection of the occurrences of a word-form, each in its own textual environment ». La concordance, telle que définie dans les années 90, est donc établie par le lexicographe pour répondre à des besoins spécifiques, notamment la délimitation des sens d'un mot, l'étude des restrictions sémantiques auxquelles ce mot est soumis, l'analyse de la complémentation ou encore l'utilisation des majuscules.

Une concordance peut prendre la forme d'un *Key-Word-In-Context*, ou KWIC, qui est devenu, dans l'usage à tout le moins, synonyme de concordance. On appelle KWIC une concordance dans laquelle « chaque mot devient une vedette, et [ce mot] est placé au milieu d'une ligne de contexte naturel » (Susanne Hanon 1990b:26). Cette définition n'est pas sans rappeler celle de Sinclair (1991:32) « the [...] KWIC format prints a whole line of text, with the word under examination in the middle ». C'est sous cette forme, dans un premier temps, qu'une concordance est le plus utile en lexicographie. Il est plus facile d'y repérer les regroupements syntaxiques et lexicaux, détails qui sont d'une importance capitale dans l'analyse d'un mot-vedette et, subséquemment, dans la rédaction de son entrée. Habituellement, on trouve aussi au début de chaque ligne de concordance une référence numérique qui permet au programme de replacer le KWIC en contexte dans le corpus. Dans l'exemple qui suit[1] (tableau 5), les chiffres en début de ligne indiquent avec

---

[1] Cet exemple est extrait de la concordance pour le mot *public* tirée du sous-corpus Presse canadienne française par le concordancier unilingue PAT. Ce logiciel a été conçu spécialement pour la lexicographie par la société Open Text de Waterloo, Ontario. C'est habituellement cet outil que les lexicographes du DCB utilisent pour consulter TEXTUM, un énorme corpus de plus de 300 millions de mots.

précision à quel caractère du sous-corpus la ligne de concordance commence puisque chaque caractère (lettre, signe de ponctuation, espace) est numéroté. Ainsi, pour la première ligne de concordance du mot *public*, le « l » au début de la ligne se situe au 64207403e caractère du sous-corpus Presse canadienne française (PCF).

Extrait de la concordance pour le mot *public*

```
>> "public "
  1: 23151 matches

>> pr sample
 64207403, ..laboration du public afin de retrouver une adolescente de 15 ans..
337304579, ..nt un service public.  Certains éditorialistes ont très bien not..
274518153, ..ntact avec le public de Québec,en 1990. L'iconoclartiste se sent..
397173341, ..rapport rendu public en octobre. Constatant une évolution de la ..
  6192800, ..es - Secteurs public et para-public - Québec.  *Grèves et lock-o..
122305955, ..pectacle tout public. Il dépayse et porte à la réflexion, et cel..
373093939, ..croissante du public, le président Bill Clinton essayait hier de..
124263130, ..s. Le secteur public ne peut échapper à cette réalité.  ILLUSTR..
270202571, ..u la santé du public pourraient être mises en péril.  Il n'y a p..
418711697, ..ulla, mais le public se demande ce que l'on attend pour corriger..
```

Tableau 5 – *Concordance pour le mot public*

Sans entrer dans une analyse détaillée de ces dix lignes de concordance (le but de cette thèse n'étant pas de montrer la pertinence des concordances monolingues en lexicographie), on remarquera d'emblée que *public* peut être un substantif ou adjectif. De plus, quelques combinaisons sautent aux yeux, soit *service public*, *secteur public* et *rendre public*, combinaisons qui doivent manifestement faire partie de l'entrée pour l'adjectif *public*.

Mais si le KWIC est le type le plus commun de concordance, il n'est pas le seul. On utilise aussi un KWOC (*Key-Word-Out-Of-Context*), qui est l'index de tous les mots contenus dans un corpus ainsi que leur fréquence (tableau 6).

Exemple de KWOCs tiré du Hansard
(extrait)

7 abaissent	1 abaisserons
1 Abaisser	3 abaisseront
281 abaisser	18 abaissis
11 abaissera	1 Abaissez
2 abaisserai	3 abaissez
1 abaisseraient	4 abaissions
3 abaisserais	1 Abaissons
8 abaisserait	7 abaissons
1 abaisserez	1 Abandon
1 abaisseriez	329 abandon
4 abaisserions	

Tableau 6 – KWOCs tirés du Hansard

Certains logiciels combinent les formats KWIC et KWOC, en affichant d'abord le mot demandé et sa fréquence, et ensuite des contextes d'une ligne dans lesquels le mot-clé apparaît. PAT, par exemple, affiche d'abord le nombre d'occurrences du mot demandé dans le corpus; il revient ensuite à l'utilisateur de décider s'il veut afficher toutes les lignes de concordance ou seulement un échantillon. Le tableau 5 illustre bien cette combinaison de KWOC, suivi de KWIC. La forme *public* suivie d'un espace revenant 23 151 fois dans le sous-corpus, l'utilisateur a jugé bon de n'en imprimer que 10 avec la commande *pr sample*.

Les concordances sont produites par un logiciel d'analyse textuelle conçu à cet effet. Ce logiciel s'appelle *concordancier* en français, mais la terminologie anglaise n'est pas encore arrêtée. En effet, selon les auteurs, on parle de *concordancer*, de *concordancing software*, de *concordancing program et de concordance generator*[2], (Rundell et Stock 1992a:10), de *concordance program* et de *concordancer* (Leech 1992:114). Mais puisque c'est le propre de toutes les langues de tenter de faire court, le terme simple *concordancer* semble avoir de plus en plus la faveur des chercheurs.

QUERY, un des premiers concordanciers utilisés sur le corpus de Brown étiqueté, avait déjà produit des résultats intéressants. Par exemple, la requête « * TO R * V » repérait tous les *split infinitives*[3]; la requête « * TO * . # », tous les *to* en fin de phrase[4] (tableau 7); et la série de caractères suivants « # * % * W - ? . # .EN. .NIET. # * % W - * V - * V - ? . # », toutes les phrases interrogatives monoverbales commençant par un pronom interrogatif en *Wh*[5].

---

[2] TRIBBLE, C. et G. JONES: *Concordances in the Classroom*, Londres, Longman, 1990. Extrait publié dans *English Today*, n° 31, pp. 29-32. Ce livre n'est malheureusement plus disponible.

[3] Un infinitif séparé de sa particule *to* était considéré, à l'époque à laquelle le corpus de Brown a été réuni, comme étant une forme fautive en anglais. Fait intéressant, sur les 25 instances de *split infinitive* trouvées dans ce corpus d'un million de mots, sept provenaient de sources gouvernementales et six du corpus scientifique, tandis qu'aucune occurrence de ce phénomène n'avait été repérée dans le corpus littéraire (Willem Meijs 1982:35).

[4] Meijs (1982:37).

[5] Meijs (1982:39).

Extrait des résultats livrés par la requête
**** TO * . #***
avec le logiciel QUERY sur le corpus de Brown[6]

---------- CORPUS A ----------

1695 - 19 I'll write what you tell me to.

---------- CORPUS F ----------

4701 - 27 I know now why the students insisted that I go to
Hiroshima even when I told them I didn't want to.

---------- CORPUS G ----------

4635 - 48 while Thomas' injured back led him to restrain his mount from
it's most violent gait he moved quickly enough when he had to.
5748 - 60 my father, a wise man, asked him not to.

Tableau 7 – QUERY : requête * *TO* * . #

Malgré sa syntaxe des plus rébarbatives, QUERY traitait quand même des requêtes assez complexes. Il présentait cependant des inconvénients évidents : il ne travaillait que très lentement et seulement sur des corpus relativement petits (si l'on compare aux corpus actuels).

Les concordanciers utilisés par l'équipe du Cobuild au tout début du projet (Jeremy Clear 1990:96) étaient encore assez rudimentaires en comparaison avec des concordanciers actuels. Leur corpus de 7,3 millions de mots a dû être divisé en six tranches, car le logiciel ne pouvait traiter tout le corpus en une seule fois. Pour faciliter encore plus la tâche de l'ordinateur, on ne

---

[6] Meijs (1982:35).

générait des concordances qu'une lettre à la fois. Et pas question de lancer des requêtes spécifiques : les concordances complètes étaient imprimées à l'avance, et les lexicographes consultaient au besoin les concordances *papier* dans de grands cahiers.

Au cours des dernières années, les concordanciers ont considérablement évolué. Maintenant, ils tournent beaucoup plus vite sur des corpus des centaines de fois plus volumineux que le corpus de Brown. Il est aussi possible de les utiliser de façon interactive : ainsi, pour obtenir la concordance pour un mot donné, il suffit de taper le mot en question; les lignes de concordance apparaissent rapidement à l'écran.

Les concordanciers fiables et performants d'aujourd'hui permettent de peaufiner les requêtes et de modifier la présentation des résultats. En effet, on peut combiner plusieurs mots dans la syntaxe d'interrogation, demander que le mot-vedette soit suivi ou non par un mot donné, etc. Les tableaux 8 et 9 serviront à illustrer la puissance des concordanciers actuels.

Extrait des résultats d'une requête combinant plusieurs mots[7]

```
>> "ami "
  1: 1178 matches

>> "ami "+"amis "+"amie "+"amies "
  2: 2332 matches

 24838330, ..lièrement son ami Pryor avec qui il n'avait jamais tourné. Et a ..
  8672275, .. suivante, un ami soviétique a réussi à m'inscrire à l'Institut ..
 24386370, ..a suggéré une amie commune, Élise Benoit, une psycho-éducatrice ..
 23305031, ..êchant son ex-amie de mettre un terme à sa grossesse.     Le 17 ..
 33307297, ..andis que son amie était étendue non loin, face contre terre dan..
 33217258, ..s soeurs, vos amies, futures ingénieures que nous pleurons aujou..
 38135855, ..es parents ou amis à Mirabel ou à Dorval au cours de la fin de s..
 12205135, ..'il avait des amis au sein du conglomérat Paramount Communicatio..
```

Tableau 8 --Requête combinant plusieurs mots

Extrait des résultats d'une requête dans laquelle on exige qu'un premier mot soit suivi par un second

```
>> pays fby ami
  1: 32 matches

 42486917, ..Cuba était un pays ami de Panama et de l'Union soviétique? Voilà..
 23631820, ..e froisser un pays ami , dit M. Roche, en entrevue à La Presse. ..
 51669678, ..conomique de  pays amis .     Joe Sanders, conseiller politique ..
 13506188, ..te l'aide des pays arabes et amis pour obtenir le départ des for..
 15077176, .. que les deux pays étaient des  amis et alliés de longue date . ..
 53924824, ..et un nouveau pays sans amis et sans parents.      Il est déçu d..
 30749182, ..it seul, sans pays, sans beaucoup d'amis intimes non plus semble..
 24162699, ..ons que notre pays soit traité en partenaire et en ami , a-t-il ..
 25510036, ..ous sommes un pays voisin, un ami, une fédération, a-t-il déclar..
```

Tableau 9 -- Requête dans laquelle un premier mot est suivi d'un second

Les concordanciers actuels permettent aussi de préciser la longueur des contextes (une phrase complète, un paragraphe complet ou un certain nombre de caractères fixé par l'utilisateur, selon

[7] Les concordances présentées dans les tableaux 8 et 9 ont été produites par le logiciel PAT sur le sous-corpus de La Presse.

le logiciel) ou de modifier la présentation des KWIC. En effet, ces derniers peuvent être classés dans l'ordre alphabétique du mot qui précède ou suit le mot-vedette, comme le montrent les tableaux 10 et 11.

Extrait de la concordance pour le mot *rédigé*
en ordre alphabétique du mot qui le suit

```
22143752, ..nt a en effet rédigé des rapports détaillés sur le sujet et sur ..
17919029, .. anglais.      Rédigé en anglais dans l'original, le guide présen..
20416377, .. années; il a rédigé le mémoire de l'ACRTF et le soutiendra deva..
 1704558, ..J'ai toujours rédigé mes contrats moi-même. Je me suis souvent m..
21870889, .. train, a été rédigé par le journaliste Gil Courtemanche.      Le..
 6886689, ..re un message rédigé par le président Jimmy Carter le 16 juillet..
43484271, ..s ce document rédigé par un individu aux prises avec des problèm..
```

Tableau 10 – *Concordance pour le mot rédigé (ordre alphabétique du mot qui suit)*

Extrait de la concordance pour le mot *rédigé*
en ordre alphabétique du mot qui le précède

```
20416377, .. années; il a rédigé le mémoire de l'ACRTF et le soutiendra deva..
17919029, .. et anglais.  Rédigé en anglais dans l'original, le guide présen..
43484271, ..s ce document rédigé par un individu aux prises avec des problèm..
22143752, ..nt a en effet rédigé des rapports détaillés sur le sujet et sur ..
21870889, .. train, a été rédigé par le journaliste Gil Courtemanche.      Le..
 6886689, ..re un message rédigé par le président Jimmy Carter le 16 juillet..
 1704558, ..J'ai toujours rédigé mes contrats moi-même. Je me suis souvent m..
```

Tableau 11 – *Concordance pour le mot rédigé (ordre alphabétique du mot qui précède)*

Le logiciel PAT dont nous avons déjà parlé n'offre malheureusement pas cette dernière possibilité, mais d'autres concordanciers comme TACT[8], eux, le font. Cette flexibilité de l'engin de

[8] TACT a été mis au point par le *Centre for Computing in the Humanities* de l'université de Toronto.

recherche est certes très utile en lexicographie, puisqu'elle permet au lexicographe de repérer des combinaisons ou des types de combinaisons qui auraient pu passer inaperçues. Par exemple, les tableaux 10 et 11 permettent de voir que le verbe *(avoir) rédigé* se met souvent au passif et ils facilitent le repérage du type de *choses* que l'on peut, en fait, rédiger.

2.2 Bi-concordance et bi-concordancier

Les concordanciers utilisés pour les corpus monolingues, aussi élaborés soient-ils, ne fonctionnent pas sur les bitextes. Il a fallu mettre au point des outils qui permettent, justement, de tirer profit de ces textes appariés. Ces *bilingual concordance (generating) programs* (Church et Gale 1991b:7) et *programmes de concordances bilingues* (Isabelle et Warwick-Armstrong 1993:299) « search bi-text to produce concordances in which a match consists of a source language segment together with its translation » (Lucie Langlois 1996:35). Dans le cadre du présent travail, les programmes en question se nommeront *bi-concordanciers* (ou *bi-concordancers* en anglais), et les résultats, des *bi-concordances*.

Church et Gale (1991a) proposent deux types de bi-concordanciers. Le premier, qu'ils appellent *sentence-based concordance program*, repère des phrases en L1 et en L2 étant donné un mot en L1 et un mot en L2. Ainsi, l'utilisateur doit savoir à l'avance ce qu'il veut étudier. L'exemple présenté par les auteurs est celui du mot *drug* en anglais : on doit lancer une requête en précisant à l'avance le couple *drug-drogue* ou *drug-médicaments*, et le programme affiche tous les couples de phrases qui répondent aux critères précisés.

L'autre type de programme, le *word-based concordance program*, extrait, pour un mot donné en L1, la concordance de ce mot et les traductions avec lesquels les passages en L1 sont appariés. Cette méthode permet de repérer des couples de phrases dans lesquels un équivalent direct du mot demandé n'apparaît pas nécessairement, puisqu'il ne faut pas le préciser au départ. Les auteurs présentent le cas du même mot anglais, soit *drug*, traité par ce type de bi-concordancier. Parmi les couples repérés, on retrouve *The drug was simply banned* et sa traduction française, *Ce dernier a été simplement interdit*, traduction qui ne comprend pas d'équivalent direct de *drug*.

Si ces deux types de bi-concordanciers peuvent faciliter la tâche du lexicographe, ce sont les possibilités booléennes de tels programmes qui permettent de raffiner les requêtes. La prochaine section présentera un bi-concordancier qui allie flexibilité de recherche, puissance de calcul et facilité d'utilisation.

2.3 Description de TransSearch

TransSearch, le bi-concordancier que nous avons utilisé dans le présent travail, a été mis au point par les chercheurs de l'équipe de Traduction assistée par ordinateur du CITI. Ce bi-concordancier permet à l'utilisateur d'imposer un nombre considérable de contraintes sur les couples de phrases qui seront repérés, ce qui en fait un outil de choix pour élargir ou raffiner des requêtes. Diverses fonctions du logiciel qui sont particulièrement intéressantes pour la lexicographie seront présentées ici, exemples à l'appui.

2.3.1 Changement de langue[9]

Il est important que l'usager puisse préciser la langue dans laquelle TransSearch doit repérer le mot demandé. Cette option est primordiale, certains mots ayant la même forme graphique en français et en anglais. Comme l'illustrent les tableaux 12 et 13 suivants, la forme *cordial*, qui peut être un mot français ou anglais, donnera des bi-concordances bien différentes selon qu'on le repère dans la partie française ou anglaise du bitexte.

Changement de langue - anglais	
Requête : cordial (anglais)	
Résultat : 14 matches in 14 couples, 100% of search done. *(extrait)*	
We all enjoyed those warm and **cordial** times in Mr. Speaker's office.	Nous avons tous apprécié ces moments chaleureux dans le bureau du Président.
Since he assumed I was in the House, he was particularly **cordial**, but unfortunately I was on the telephone in the lobby, trying to clear up an immigration problem, so I did not hear a word of what he said.	Croyant que j'étais à la Chambre, il en avait mis un peu plus, mais, malheureusement, j'étais en train d'essayer de résoudre un problème d'immigration au téléphone dans le lobby, ce qui fait que je n'ai rien entendu de ce qu'il a dit.
I would like to respond to the **cordial** remarks made by the hon. member for South West Nova and stress the superb co-operation between our two ridings, as a result of which residents of both ridings will be visiting each other this summer to create a brotherhood of Canadians.	Je pense que la Chambre me permettra de répondre à l'excellent commentaire de l'honorable députée de South West Nova pour souligner l'excellente collaboration qui fait que les citoyens des deux comtés se visitent au cours de l'été pour bâtir cette fraternité canadienne.

[9] Les noms des diverses fonctions sont ceux qui leur ont été attribués par le CITI.

I am sure hon. members will join me in thanking the organizers and sponsors of this event, and wish our young Canadian skaters good luck and extend a **cordial** welcome to all participants.	Je sais que les autres députés se joindront à moi pour remercier les organisateurs et les commanditaires de l'événement, pour souhaiter bonne chance à nos jeunes patineurs canadiens et pour accueillir chaleureusement tous les athlètes qui participent aux championnats.

Tableau 12 – *TransSearch : changement de langue - anglais (cordial)*

Changement de langue - français	
Requête : cordial (français)	
Résultat : 2 matches in 2 couples, 100% of search done.	
We must each play a part to ensure that we provide a genuine welcome to those who have chosen Canada as their new home.	Les Canadiens doivent offrir un accueil **cordial** à es gens qui ont choisi le Canada comme nouvelle patrie.
Mr. Simmons: The short time I spent in that assignment was enjoyable all the more so because the minister was always very cordial.	M. Simmons : Le peu de temps que j'ai passé dans ces fonctions a été d'autant plus agréable que mon collègue a toujours été très **cordial**.

Tableau 13 – *TransSearch : changement de langue - français (cordial)*

2.3.2 Langue source

L'utilisateur peut aussi préciser la langue de départ dans les couples repérés, ce qui est aussi très pratique dans le contexte de la lexicographie bilingue. Même si un texte et sa traduction sont tous deux des textes d'une excellente qualité sur la plan de la langue, il est essentiel, en lexicographie bilingue, de connaître la langue de départ, d'une part pour juger de la qualité de la traduction,

d'autre part pour savoir si un locuteur natif se servirait d'un mot ou d'une expression donnée ou si c'est un mot ou expression privilégiée des traducteurs.

Examinons le cas du mot chefferie. Si TransSearch repère tous les cas de *chefferie* quelle que soit la langue source, nous obtenons 42 résultats (tableau 14). Néanmoins, si on lance la même requête en précisant cette fois que la langue source doit être l'anglais, TransSearch ne trouve aucun résultat (tableau 15), ce qui veut dire que le mot *chefferie*, dans le contexte du Hansard, n'est utilisé que par les députés francophones à la Chambre des Communes et jamais par les traducteurs.

Langue source	
Requête : chefferie (français) lorsque la langue source est indéterminée	
Résultat : 42 matches in 41 couples, 100% of search done. *(extrait)*	
I also would like to congratulate the four other candidates for the leadership of the Progressive Conservative Party.	Je voudrais également féliciter les quatre autres candidats à la **chefferie** du Parti conservateur.
I also want to pay special tribute to the hon. member for Sherbrooke for the excellent way in which he conducted his campaign.	J'aimerais également rendre un super hommage au député de Sherbrooke, l'honorable Jean Charest, pour l'excellente campagne qu'il a menée à la **chefferie** du Parti progressiste-conservateur du Canada.
All Manitobans are winners in this leadership race.	Tous les Manitobains sortent gagnants de cette course à la **chefferie**.
I was also here when both Tory leadership candidates fully supported the policies of that same minister.	J'étais ici aussi lorsque les deux candidats à la **chefferie** du Parti conservateur ont appuyé entièrement les politiques de ce même ministre des Finances.

Don't the two leadership contenders realize that in Canada health is a right, not a privilege?	Les deux aspirants à la **chefferie** ne comprennent-ils pas qu'au Canada, la santé est un droit et non un privilège?
It is an insult to our intelligence when candidates for the leadership of the Conservative Party tell us that they are concerned about the deficit.	C'est une insulte à notre intelligence quand les candidats à la **chefferie** du Parti conservateur nous disent qu'ils sont préoccupés par le déficit.
Because of the coming leadership convention the budget contains of course a lot of wishful thinking to the effect that the future is rosy and that things will get better.	À cause de la course à la **chefferie**, le budget contenait évidemment beaucoup de voeux pieux pour dire que l'avenir sera rose et que les choses s'arrangeront.

Tableau 14 – *TransSearch – langue source indéterminée (chefferie)*

Langue source
Requête : chefferie (français) lorsque la langue source est l'anglais
Résultat : 0 matche in 0 couple, 100% of search done.

Tableau 15 – *TransSearch – langue source précisée (chefferie)*

2.3.3 Création de plusieurs expressions (conjonction)

TransSearch offre la possibilité de préciser plus d'une restriction, que ce soit sur le corpus français, le corpus anglais ou les deux. Par exemple, on peut demander que le mot X apparaisse dans la partie anglaise du couple de phrases et le mot Y, dans la partie française. Cette fonction est l'équivalent du *AND* en logique booléenne. Puisque les résultats devront répondre à toutes les conditions énumérées dans la requête, le lexicographe peut ainsi contrer le trop-plein

d'information, confirmer une intuition personnelle ou encore vérifier si les équivalents proposés par les dictionnaires bilingues sont vraiment utilisés.

Dans l'exemple qui suit (tableau 16), le lexicographe a pu vérifier si, effectivement, *chefferie* était équivalent du mot *leadership* en anglais en demandant à TransSearch de repérer tous les couples de phrases dans lesquels *leadership* apparaissait du côté anglais et *chefferie* du côté français.

Création de plusieurs expressions (conjonction)	
Requête : leadership (anglais) AND chefferie (français)	
Résultat : 39 matches in 39 couples, 100% of search done. *(extrait)*	
I also would like to congratulate the four other candidates for the **leadership** of the Progressive Conservative Party .	Je voudrais également féliciter les quatre autres candidats à la **chefferie** du Parti conservateur .
All Manitobans are winners in this **leadership** race .	Tous les Manitobains sortent gagnants de cette course à la **chefferie** .
I was also here when both Tory **leadership** candidates fully supported the policies of that same minister .	J' étais ici aussi lorsque les deux candidats à la **chefferie** du Parti conservateur ont appuyé entièrement les politiques de ce même ministre des Finances .
Don't the two **leadership** contenders realize that in Canada health is a right , not a privilege ?	Les deux aspirants à la **chefferie** ne comprennent - ils pas qu' au Canada , la santé est un droit et non un privilège ?

Tableau 16 – *TransSearch – conjonction (leadership AND chefferie)*

2.3.4 Négation d'une requête

La fonction *négation* permet de nier une requête; ainsi, le bi-concordancier n'affichera que les couples qui ne répondent pas à la contrainte faisant l'objet de la négation. Elle sert surtout lorsque le lexicographe veut obtenir, pour un mot donné, des équivalents autres que les équivalents les plus évidents. La négation d'équivalents est aussi une excellente façon de limiter le nombre de résultats à analyser. Dans l'exemple qui suit (tableau 17), le lexicographe a vérifié si l'on pouvait rendre *chefferie* par un autre mot que *leadership*.

Négation d'une requête	
Requête : chefferie (français) *NOT* leadership (anglais)	
Résultat : 2 matches in 2 couples, 100.00 % of search done.	
I also want to pay special tribute to the hon. member for Sherbrooke for the excellent way in which he conducted his campaign.	J'aimerais également rendre un super hommage au député de Sherbrooke, l'honorable Jean Charest, pour l'excellente campagne qu'il a menée à la **chefferie** du Parti progressiste-conservateur du Canada.
Maybe we will see them do a `` flip flop'' , like some candidates to the direction of the party are doing regarding their position on the Meech Lake Accord. One day, they are against the agreement as it stands, but toward the end of the race, they start mellowing.	Peut-être allons-nous assister à un «flip flop», comme certains aspirants à la **chefferie** agissent actuellement dans leurs idées sur l'Accord du lac Meech; une journée, ils sont contre l'Accord du lac Meech tel quel, mais quand arrive la fin, là, ils commencent à être pour.

Tableau 17 – *TransSearch – négation (chefferie NOT leadership)*

Dans le premier couple, l'idée de *chefferie* n'est pas rendue dans la phrase anglaise tandis que, dans le deuxième couple, l'idée est rendue par *direction of the party*.

2.3.5 Création de formes alternatives (disjonction)

Cette fonction permet au lexicographe de combiner plusieurs recherches. Équivalente au *OR* booléen, elle est moins utile en lexicographie, puisque le lexicographe a plutôt tendance à vouloir restreindre et préciser ses recherches, et non à les élargir. Néanmoins, cette fonction sert notamment à trouver, sous toutes ses expansions morphologiques, un mot qui n'est pas dans le dictionnaire dont se sert TransSearch[10] ou encore à récupérer différentes variantes orthographiques d'un mot. Le tableau 18 montre les résultat d'une requête visant à connaître l'équivalent anglais du mot *cégep*, que celui-ci soit écrit en majuscules ou pas[11].

Création de formes alternatives (disjonction)
Requête : Cégep (français) *OR* CEGEP (français)
Résultat : 8 matches in 8 couples, 100.00% of search done.

[10] À titre d'exemple, nous pouvons citer le verbe taponner. Puisqu'il n'apparaît pas dans le dictionnaire morphologique qu'utilise TransSearch, le lexicographe a dû lancer la requête suivante : taponner OU taponné OU taponne OU taponnes OU taponnons OU taponnez OU taponnent OU taponnais OU taponnait, etc. pour récupérer toutes les occurrences de ce verbe.

[11] Lorsque seule la présence d'une majuscule à la première lettre d'un mot fait la différence, comme ce serait le cas pour *état* et *État*, le lexicographe également peut utiliser la fonction *Pertinence des majuscules*, expliquée ici à la sous-section 2.3.6.

Thanks to a federal contribution of $200,000, the Saint-Jérôme CEGEP will be able to add a co-ooperative education component in its curriculum.	Ainsi, la contribution de 200 000 $ du gouvernement fédéral permettra au Cégep de Saint-Jérôme d'introduire une composante travail-études dans ses programmes.
The CEGEP, in co-operation with business, will then be able to give students the tools they need to enter the labour market successfully.	Le Cégep de Saint-Jérôme, de concert avec les entreprises, sera alors en mesure de fournir aux étudiants et aux étudiantes les outils nécessaires pour mieux les intégrer au marché du travail.
Three major partners in the region have decided to create an institute for international trade, to be located at Chaudière-Appalaches. I am referring to the network for business expansion, the Beauce economic council and the Beauce- Appalaches CEGEP.	Trois importants partenaires beaucerons se sont associés pour créer l'Institut du commerce international sur le territoire Chaudière-Appalaches : Il s'agit du Réseau pour l'expansion des entreprises, le Conseil économique de Beauce et le Cégep Beauce-Appalaches.
Mr. Jacques Vien (Laurentides): Mr. Speaker, on October 26, I had the honour of announcing, on behalf of the Hon. Marcel Danis, that a co-operative education project will start at the Saint-Jérôme CEGEP.	M. Jacques Vien (Laurentides) : Monsieur le Président, le 26 octobre dernier, j'avais l'honneur d'annoncer, au nom de l'honorable Marcel Danis, la mise en oeuvre d'un projet Alternance travail- études au Cégep de Saint-Jérôme.
The Saint-Jérôme CEGEP is to be commended for its involvement in the community, and I wish those students who take advantage of the co-operative education option every success.	Je félicite le Cégep de Saint-Jérôme pour son implication au sein de la communauté et je souhaite aux étudiants et aux étudiantes qui bénéficieront de l'option travail-études les meilleures chances de succès.
The CEGEP, in co-operation with business, will then be able to give students the tools they need to enter the labour market successfully.	Le Cégep de Saint-Jérôme, de concert avec les entreprises, sera alors en mesure de fournir aux étudiants et aux étudiantes les outils nécessaires pour mieux les intégrer au marché du travail.
The Saint-Jérôme CEGEP is to be commended for its involvement in the community, and I wish those students who take advantage of the co-operative education option every success.	Je félicite le Cégep de Saint-Jérôme pour son implication au sein de la communauté et je souhaite aux étudiants et aux étudiantes qui bénéficieront de l'option travail-études les meilleures chances de succès.
Then Quebec introduced the CEGIP system which also was a quasi post-secondary educational institution.	Puis le Québec a créé les CEGEP, qui étaient une forme d'établissement d'enseignement postsecondaire.

Tableau 18 – *TransSearch – disjonction (Cégep OR CEGEP)*

2.3.6 Pertinence des majuscules

La fonction *Pertinence des majuscules* permet au lexicographe de préciser si, oui ou non, il veut utiliser la présence d'une majuscule comme critère de sélection. Si l'usage des majuscules peut changer avec le temps, il n'en reste pas moins que celles-ci sont très importantes en lexicographie descriptive; un dictionnaire doit en tenir compte. De plus, il peut être utile, par exemple, de pouvoir éliminer toutes les instances des noms propres *Fisher* et *Fish*, si les noms communs *fisher* et *fish* sont à l'étude, ou encore de voir si un certain adverbe, comme *however*, est souvent utilisée en début de phrase. Dans le cas de *cégep* discuté plus haut, la fonction *Pertinence des majuscules* aurait facilité la tâche du lexicographe s'il avait voulu vérifier la fréquence des différentes variantes orthographiques (*cégep, Cégep, CEGEP* et *CÉGEP*).

2.3.7 Type de recherche

Le bi-concordancier TransSearch est assorti d'un vaste dictionnaire morphologique; ainsi, l'utilisateur a le choix entre une recherche *exacte* ou une recherche *dictionnaire*. La première option ne récupère que la chaîne de caractères demandée tandis que la recherche *dictionnaire* donne la fréquence totale d'un mot ou d'une expression sous toutes ses formes morphologiques sans que le lexicographe ait à lancer des requêtes pour toutes les variantes possibles. Le prochain exemple (tableaux 19 et 20) montre la différence entre ces deux types de recherche. Dans le premier cas (tableau 19), TransSearch n'a récupéré que la chaîne de caractères *mettre le doigt dans l'oeil* (une seule occurrence). Le tableau 20 montre que TransSearch a aussi trouvé les variantes

met/mettent/mettre le doigt dans l'oeil lorsqu'on lui a demandé d'effectuer une recherche dictionnaire pour cette même expression.

Recherche exacte	
Requête : mettre le doigt dans l'oeil (français)	
Résultat : 1 match in 1 couples, 100.00% of search done.	
He would do well not to put his foot so far down his throat as he has today ever again. Never again. Because he insults every Newfoundland-born member in his constituency, every person in P.E.I. and so many other provinces, and insults his colleagues from ridings like Broadview--Greenwood and Burlington.	Il ferait mieux dorénavant de ne pas se **mettre le doigt dans l' oeil** aussi profondément, parce qu'il insulte tous les électeurs dans sa circonscription qui sont originaires de Terre-Neuve, tous les Terre-Neuviens, bien des provinces et ses collègues des circonscriptions comme Broadview--Greenwood et Burlington. Certaines personnes ne se rendent pas compte qu'ils sont la cible de ses injures.

Tableau 19 – *TransSearch – recherche exacte (mettre le doigt dans l'oeil)*

Recherche dictionnaire	
Requête : mettre(+) le doigt dans l'oeil (français)	
Résultat : 3 match in 3 couples, 100.00% of search done.	
I will not give this thing any more importance than to say, you are all wet.	Je n'attacherai pas plus d'importance à cette affaire que de dire aux ministériels qu'ils se **mettent le doigt dans l' oeil.**
If this government thinks that it is going to be able to have as much control over the breeding, proliferation and even the transportation of plants, cereal grains and so on across the border as it has now, then it is certainly very much mistaken.	Si le gouvernement croit pouvoir garder le contrôle qu'il exerce actuellement sur la création, la production, voire le transport transfrontalier des plantes, céréales et autres végétaux, il se **met le doigt dans l' oeil.**
He would do well not to put his foot so far down his throat as he has today ever again. Never again. Because he insults every Newfoundland-born member in his constituency, every person in P.E.I. and so many other provinces, and insults his colleagues from ridings like Broadview--Greenwood and Burlington.	Il ferait mieux dorénavant de ne pas se **mettre le doigt dans l' oeil** aussi profondément, parce qu'il insulte tous les électeurs dans sa circonscription qui sont originaires de Terre-Neuve, tous les Terre-Neuviens, bien des provinces et ses collègues des circonscriptions comme Broadview--Greenwood et Burlington. Certaines personnes ne se rendent pas compte qu'ils sont la cible de ses injures.

Tableau 20 – *TransSearch – recherche dictionnaire (mettre le doigt dans l'oeil)*

2.3.8 Distance

Les mots qui forment des collocations n'apparaissent pas nécessairement l'un immédiatement après l'autre. C'est pourquoi, avec la fonction *distance*, l'utilisateur peut préciser la distance maximale qui peut séparer deux mots, ce qui lui permet de juger de la flexibilité de ces combinaisons. Si, pour repérer les mots A et B, on ne précise pas une distance maximale, les résultats incluront toutes les phrases qui comprennent et le mot A et le mot B, nonobstant la distance qui les sépare.

Il va de soi que plus la distance entre ces mots est grande, moins fortes sont les chances que A et B forment un syntagme[12].

Ainsi, il arrive que le lexicographe joue avec ce paramètre pour trouver la distance optimale entre deux mots dans le but de récupérer un nombre maximum d'instances sans pour autant produire trop de bruit. Dans l'exemple présenté ici (tableaux 21 et 22), TransSearch a récupéré *pousser* et *cri* sous toutes leurs formes morphologiques. Dans le premier cas (tableau 21), ils ne sont séparés l'un de l'autre que par un mot, dans l'autre cas, par 3 mots. Quatre cas de *pousse(+) des cris* ont été repérés par la première requête. Cependant, en faisant passer la distance à 3 (tableau 22), le lexicographe a récupéré 18 couples, dont 14 étaient des instances de *pousser(+) les/des hauts cris*, une variante dont le dictionnaire devra sûrement tenir compte.

Mots non consécutifs - Distance de 1	
Requête : pousser(+) (français) *[distance = 1]* cri(+) (français)	
Résultat : 4 match in 4 couples, 100.00% of search done.	
The hon. member would be on his feet righteously screaming in indignation if the post office sent out a memo saying it is okay to backdate the stamp on income taxes so if one's friend brings in his income tax late it is all right to backdate it.	Le député aurait eu tout à fait raison de **pousser des cris** d'indignation si Postes Canada avait envoyé une note pour demander à ses employés d'antidater les déclarations d'impôt. Ainsi, si quelqu'un apportait sa déclaration en retard, il n'y aurait rien de mal à antidater l'enveloppe.
Because the hon. member rises in this House and shouts across the House does not make a situation a crisis.	Ce n'est pas parce que le député se lève à la Chambre et **pousse des cris** qu'il y a une situation de crise.

[12] La plupart des chercheurs estiment que la distance maximale qui peut séparer une base de son collocatif est 5 mots.

I said to the hon. member: `` Because the hon. member rises in this House and shouts across the House does not make the situation a crisis.	J'ai dit au député : «Ce n'est pas parce que le député se lève à la Chambre et **pousse des cris** qu'il y a une situation de crise.
`` The people can't stand your new tax! It's regressive, oppressive and unfair! So be a good boy and join with me! As together, we all declare''	Parce que c'est de l'oppression! C'est de l'injustice, c'est régressif! Souris, voyons, sois bon garçon! Avec tout le monde **pousse ce cri** expressif :

Tableau 21 – *TransSearch – distance de 1 (pousser + cri)*

Mots non consécutifs - Distance de 3	
Requête : pousser(+) (français) *[distance = 3]* cri(+) (français)	
Résultat : 18 match in 18 couples, 100.00 % of search done. *(extrait)*	
The hon. member should consider that before he gets indignant.	Le député devrait réfléchir à cela avant de **pousser les hauts cris.**
The hon. member would be on his feet righteously screaming in indignation if the post office sent out a memo saying it is okay to backdate the stamp on income taxes so if one's friend brings in his income tax late it is all right to backdate it.	Le député aurait eu tout à fait raison de **pousser des cris** d'indignation si Postes Canada avait envoyé une note pour demander à ses employés d'antidater les déclarations d'impôt. Ainsi, si quelqu'un apportait sa déclaration en retard, il n'y aurait rien de mal à antidater l'enveloppe.
Whenever we suggest that, we get cries from the other side that somehow we are intervening and we are playing in the marketplace and we should not.	Chaque fois que l'on dit cela, il y en a de l'autre côté qui **poussent les hauts cris**, sous prétexte que ce serait de l'ingérence dans le marché, ce que l'on doit éviter à tout prix.
Because the hon. member rises in this House and shouts across the House does not make a situation a crisis.	Ce n'est pas parce que le député se lève à la Chambre et **pousse des cris** qu'il y a une situation de crise.
Canadians have complained loudly and bitterly but they re-elected him in 1988. He has kept the squeeze on, although he will have to run again next year.	Les Canadiens **ont poussé les hauts cris** et s'en sont plaints amèrement, mais ont quand même, en 1988, réélu M. Mulroney qui continue d'imposer ces mesures d'austérité, bien qu'il ait l'intention de se représenter aux élections l'an prochain.

Tableau 22 – *TransSearch – distance de 3 (pousser + cri)*

2.3.9 Examen du document

Il est possible, avec la fonction *Examen du document*, de remettre en contexte un couple de phrases affiché en concordance. Si elle n'est pas souvent utilisée en lexicographie, on s'en sert parfois pour mieux comprendre une traduction donnée.

2.3.10 Affichage du nombre d'occurrences

Si le lexicographe s'intéresse surtout à l'aspect *qualitatif* des concordances, il ne doit pas pour autant négliger le côté *quantitatif* de la langue. Il doit savoir exactement combien de couples répondent aux restrictions imposées dans sa requête : combien de fois le mot A et le mot B apparaissent dans les parties française et anglaises du bitexte respectivement ou, à l'inverse, combien de fois le mot B ne figure pas dans les phrases alignées avec celles contenant le mot A, ou encore quelle variante orthographique est la plus utilisée. Le nombre d'occurrences (ou *matches* dans les termes du logiciel) apparaît au bas de la fenêtre des résultats suivi du pourcentage du corpus qui a été balayé.

2.4 Utilisation des bi-concordances

Les bi-concordances que produisent les bi-concordanciers se sont avérées particulièrement utiles en lexicographie bilingue, comme en fait foi l'expérience menée au Projet de dictionnaire canadien bilingue. Les exemples qui suivent tentent de montrer de quelle façon les bi-concordances,

produites ici par TransSearch, peuvent être utilisées, mais la liste d'applications possibles est loin d'être exhaustive.

2.4.1 Augmentation du nombre d'équivalents présentés dans les dictionnaires

Tout traducteur, aussi chevronné soit-il, doit occasionnellement avoir recours à un dictionnaire bilingue. Et combien de fois le remet-il sur la tablette, sans y avoir trouvé de réponse? Il arrive que le dictionnaire ne propose qu'un équivalent, l'équivalent le plus évident, sans ouvrir la porte à d'autres possibilités.

C'est le cas, par exemple, de l'adverbe *automatically* cherché dans trois dictionnaires bilingues anglais/français[13]. Tandis que tous les trois proposent *automatiquement* comme équivalent, un seul mentionne *machinalement* comme autre possibilité, sans précision aucune quant à son emploi. Il faut mentionner, néanmoins, que *d'office* est utilisée dans la traduction d'une combinaison libre et que deux dictionnaires ont inclus le terme juridique *automatically void* (nul de plein droit) à la fin de l'entrée.

Il serait tentant de rédiger l'entrée pour *automatically* sur le même modèle. Cependant, une recherche dans le Hansard apparié a révélé d'autres possibilités de traduction. En fait, lorsque le lexicographe a éliminé toutes les paires de phrases qui contenaient *automatiquement* en français,

---

[13] Dictionnaire Hachette-Oxford, Grand Dictionnaire Larousse, Dictionnaire Robert & Collins Senior.

automatically était traduit par autre chose qu'*automatiquement* dans 25 p. 100 des cas, comme le montre le tableau 23.

AUTOMATICALLY	
Requête TransSearch	**Nombre de couples**
automatically (anglais)	270
automatically (anglais) **NOT** *automatiquement* (français)	82

Tableau 23 – *TransSearch – automatically NOT automatiquement*

Une analyse des couples dans lesquels *automatiquement* ne figurait pas montre diverses possibilités de traduction qui pourraient facilement être incluses dans une entrée de dictionnaire bilingue. Par exemple, il arrive que le sens de l'adverbe anglais soit rendu par l'adjectif français équivalent, ce que Jean-Paul Vinay et Jean Darbelnet (1958:96) appellent une transposition :

to pay dues automatically:
des contributions automatiques

Bien entendu, ceci ne veut pas dire que l'adjectif *automatique* devrait figurer parmi les équivalents de *automatically*. Il pourrait cependant être donné comme possibilité de traduction dans une

combinaison libre. La bi-concordance a aussi montré qu'on utilise parfois la locution adverbiale *de façon automatique* pour traduire *automatically*. Ainsi, l'entrée pour *automatically* pourrait inclure l'exemple suivant dans la section des combinaisons libres :

> *to pay dues automatically:*
> contribuer automatiquement = payer les primes par versement automatique = verser les primes de façon automatique

Les traducteurs du Hansard utilisent parfois des synonymes partiels d'*automatiquement*, comme *forcément* et *nécessairement*. Ces adverbes ne peuvent pas toujours se substituer à *automatiquement*, et il ne faut surtout pas induire l'utilisateur d'un dictionnaire en erreur. Néanmoins, un exemple bien choisi, comme celui qui est présenté ci-dessous, peut donner à l'utilisateur une bonne idée des contextes dans lesquels ces synonymes partiels peuvent remplacer *automatiquement*.

> *when the economy is weak, government automatically takes in less in taxes :*
> lorsque l'économie est au ralenti, le gouvernement perçoit automatiquement/nécessairement/forcément moins de taxes

S'il est vrai qu'un traducteur de métier traduira souvent *automatically* par autre chose qu'*automatiquement*, ce n'est pas toujours le cas pour l'apprenti-traducteur ou pour celui qui traduit hors de sa langue maternelle. En conséquence, il est intéressant de présenter des exemples plus variés et élaborés dans un dictionnaire bilingue pour diriger les utilisateurs vers la myriade de possibilités existant dans la langue d'arrivée et, plus important sans doute, pour encourager les utilisateurs à penser, eux aussi, à des façons originales de traduire un mot en contexte.

2.4.2 Confirmation d'un équivalent proposé dans les dictionnaires

Il arrive, en lexicographie bilingue, qu'un équivalent trouvé dans les dictionnaires semble douteux, soit parce qu'il est tout simplement faux ou encore parce que le lexicographe n'est pas tout à fait sûr du sens du mot ou syntagme de départ. Ici encore, une bi-concordance peut être particulièrement utile, comme le montrera l'exemple de *pie in the sky*.

Peu de francophones connaissent l'expression *pie in the sky*. Ainsi, un lexicographe francophone aux prises avec cette expression n'est peut-être pas dans une position idéale pour évaluer la pertinence des équivalents proposés dans les trois dictionnaires déjà cités (*ce sont des promesses en l'air, ce sont de belles promesses/paroles* et *c'est de l'utopie*).

Grâce à la bi-concordance, il est maintenant facile de confirmer que, effectivement, *belles promesses* et *promesses en l'air* sont des traductions de *pie in the sky*. De plus, puisque les

couples affichés sont des phrases complètes, le lexicographe peut remarquer que ces expressions équivalentes ne s'insèrent pas toujours dans des structures semblables :

> *they are not looking for pie in the sky :*
> ils ne veulent pas qu'on les abreuve de belles promesses
>
> *what we have seen in this budget is more pie in the sky :*
> ce budget contient encore d'autres belles promesses en l'air

Ainsi, non seulement la bi-concordance a-t-elle permis de confirmer les équivalents proposés par les dictionnaires bilingues, mais le lexicographe a aussi pu les voir en contexte, ce qui est particulièrement important puisque les expressions en question ne s'utilisent pas de la même façon, et l'entrée ou la sous-entrée pour *pie in the sky* devra en tenir compte.

2.4.3 Traduction de collocations

Il est maintenant essentiel d'inclure le plus grand nombre de collocations possibles dans un dictionnaire, surtout dans un dictionnaire bilingue, puisque ces faits de langue se traduisent rarement mot à mot[14].

---

[14] La collocation sera traitée à fond au Chapitre 3.

La collocation *to make history*, par exemple, pose problème. Deux des dictionnaires pré-cités donnent *entrer dans l'histoire* comme équivalent, le troisième, *être historique*. Il est bien évident que cette dernière traduction ne pourrait pas être utilisée dans tous les cas pour traduire *to make history*. Un bon traducteur ne traduirait jamais *Today, we made history* par *Aujourd'hui, nous avons été historiques*!

La bi-concordance produite par TransSearch montre bien que *entrer dans l'histoire* est une façon idiomatique de traduire *to make history*. De plus, un éventail de traductions possibles sont aussi présentées, comme *créer un précédent*, *pour la première fois dans l'histoire* et *écrire une page d'histoire*. Pour se convaincre de la pertinence de ces traductions, le lexicographe pourra ensuite consulter un corpus monolingue et inclure dans l'entrée celles qu'il jugera les plus pertinentes.

2.4.4 Traduction d'expressions figées

Les expressions figées sont particulièrement difficiles à traduire puisque, d'une part, elles ne se traduisent pas mot à mot et que, d'autre part, on ne peut les paraphraser sans une perte quant à l'image créée par l'expression originale. Elles diffèrent des collocations en ce qu'elles sont particulièrement rigides quant à leur structure. Souvent, les dictionnaires bilingues n'incluent que peu d'expressions figées, peut-être à cause de leur nature complexe. Ainsi, le traducteur qui doit traduire une expression qu'il ne connaît pas risque de perdre beaucoup de temps à chercher çà et là pour trouver une solution à son problème. Idéalement, un dictionnaire bilingue devrait fournir,

pour chaque expression figée dans une langue, une expression figée dans l'autre langue plutôt qu'une simple explication.

Le cas de *to jump/climb on the bandwagon* en est un bon exemple. Les trois dictionnaires bilingues pré-cités s'entendent quant à l'expression équivalente, car tous trois donnent *prendre le train en marche*. Un dictionnaire donne aussi *suivre le mouvement* et un autre, *se mettre dans le mouvement*. Ces deux derniers équivalents, s'ils ne constituent pas des expressions figées, expliquent néanmoins l'expression anglaise du départ. La recherche du mot mot *bandwagon* avec TransSearch a fourni d'autres équivalents intéressants, comme l'atteste le tableau 24.

Traduction d'une expression	
Requête : bandwagon (anglais)	
Résultat : 14 matches in 13 couples, 100.00% of search done.	
We have Senators Henry Waxman and David Pryor suggesting that Canada has found ways to control prescription drug prices and it is time for the United States to get on the **bandwagon.**	Les sénateurs Henry Waxman et David Pryor soutiennent en effet que le Canada a trouvé des moyens de freiner les prix des médicaments et qu'il est temps que les États-Unis lui emboîtent le pas.
It is not going to wait for Canada to get on the **bandwagon.**	Il ne va pas attendre que le Canada emboîte le pas.
We have not succumbed to the political and media pressure to just jump on any **bandwagon.**	Je puis vous assurer que nous n'avons pas cédé aux pressions politiques ni aux pressions des médias pour entrer dans la ronde comme tout le monde.
Subsequently, many others who often speak up for civil liberties in Canada, freedom of expression and so on, were latecomers to climb on board that **bandwagon.**	Par la suite, beaucoup de défenseurs des libertés civiles, de la liberté d'expression, etc., se sont joints tardivement au mouvement.
Then, he climbed on the **bandwagon**, and as soon as he was on it, he tried to push John Turner off the **bandwagon.**	Puis, il a pris le train en marche et dès qu'il y est monté, il a tenté d'en faire descendre John Turner.

At the same time, as part of this new regime of competitiveness brought to us by the Conservatives we have ended up having all those regional carriers that emerged in the rush to get in on the **bandwagon** to make a buck being acquired in some cases 100 per cent, in some cases 50 per cent and other ranges by the two major carriers.	Par ailleurs, en grande partie grâce à ce nouveau régime de compétitivité que nous ont donné les conservateurs, tous les transporteurs régionaux qui avaient surgi au moment de la déréglementation dans l'espoir de faire de l'argent, ont été acquis parfois totalement, parfois partiellement, par les deux grands transporteurs.
The accommodationists suggest that those who run the world are increasingly designing a world in which this is the stark reality and that the sooner we get on the **bandwagon** and redesign ourselves to fit this new world, the better.	Selon les partisans de l'adaptation, les dirigeants de ce monde sont en train de créer un climat où s'adapter est de pure nécessité et, plus vite nous le ferons, mieux nous nous en porterons.
That is why the New Democratic Party decided it better jump on the **bandwagon**.	C'est pour cette raison que le Nouveau Parti démocratique a jugé préférable de suivre le mouvement.
The Prime Minister has taken the easy way out. He jumped on the **bandwagon**.	Le premier ministre a choisi la voie facile en se joignant aux autres.
By jumping on the **bandwagon** called the multinational military force, Canada has lost that golden opportunity to be respected, to be counted on as the initiator of independent, thoughtful diplomacy.	En embrassant l'idée d'une force militaire multinationale, le Canada a perdu une occasion en or de se faire respecter, de s'afficher comme l'auteur d'une diplomatie indépendante et bien réfléchie.
The whole nature of what we are hearing from members of the Liberal party is that they are on a **bandwagon**.	On constate à les entendre qu'ils ont tous flairé là une bonne affaire.
Mr. Wilson (Etobicoke Centre): They are trying to jump on the **bandwagon**.	M. Wilson (Etobicoke-Centre) : Ils adoptent l'attitude la plus susceptible de leur obtenir la victoire.
They are trying to get on the **bandwagon** of opposition, saying: `` We are against this tax'' .	À l'instar des députés de l'opposition, ils disent qu'ils sont contre cette taxe.

Tableau 24 – *TransSearch – expression figée avec bandwagon*

En fait, sur 14 occurrences du mot, l'expression *prendre le train en marche* n'a été utilisée qu'une fois dans la partie française du bitexte. D'autres expressions ont été utilisées pour rendre l'idée de « s'associer à une action déjà en cours »[15], soit *emboîter le pas* et *entrer dans la ronde*, qui

[15] Définition de *prendre le train en marche* selon le Nouveau Petit Robert.

pourraient, si elles sont utilisées dans de bons exemples, enrichir grandement l'entrée pour *bandwagon*.

2.4.5 Recherche des canadianismes

Même lorsque les dictionnaires bilingues existants donnent bon nombre d'équivalents acceptables pour un mot donné, ceux-ci sont rarement des canadianismes[16]. Puisqu'il est essentiel de les inclure dans un dictionnaire bilingue *canadien*, le Hansard, de par son origine, est une source précieuse de renseignements à ce chapitre. L'exemple du mot *leadership* illustre bien cette utilité.

Lorsqu'un Canadien-français cherche le mot anglais *leadership* dans les trois dictionnaires bilingues pré-cités, il remarque une omission importante. Les équivalents proposés comprennent *direction*, *leadership* et *dirigeants*, qui sont tous des équivalents corrects de *leadership*, mais on ne mentionne aucunement le mot *chefferie*. Est-ce que *chefferie* est un équivalent canadien acceptable de *leadership*?

Trente-neuf couples ont été repérés dans le Hansard en limitant la recherche aux seules paires de phrases qui comprenaient *leadership* en anglais et *chefferie* en français (tableau 25), et tout porte à croire qu'il s'agit là d'un canadianisme.

---

[16] On entend par canadianisme un mot qui n'est utilisé qu'au Canada ou encore qui est utilisé au Canada dans un sens particulier.

LEADERSHIP	
Requête TransSearch	**Nombre de couples**
leadership (anglais)[17]	2 834
chefferie (français)	41
leadership (anglais) **AND** *chefferie* (français)	39

Tableau 25 – *TransSearch – canadianismes – leadership/chefferie*

Il va de soi que le fait de trouver quelque chose de particulier dans un corpus canadien comme le Hansard ne veut pas nécessairement dire que l'unité lexicale en question constitue un canadianisme. Cette particularité doit ensuite être confirmée dans un corpus unilingue canadien, puis infirmé dans un corpus unilingue non-canadien, avant que l'on puisse affirmer avec certitude qu'il s'agit bien là d'un canadianisme.

2.5 Utilité des bi-concordances

Nous avons tenté de montrer comment certaines fonctions de TransSearch sont utiles en lexicographie bilingue. Bien qu'au départ ce logiciel n'ait pas été conçu pour la lexicographie, il fonctionne tout de même très bien dans le contexte d'un projet de dictionnaire bilingue et il fait maintenant partie intégrante du poste de travail du lexicographe au Projet de dictionnaire canadien bilingue.

L'utilité d'un bi-concordancier comme TransSearch dépasse le simple cadre de la recherche d'équivalents et de la confirmation d'équivalents. En effet, il permet d'accéder facilement à toute

[17] Il ne faut pas s'étonner de la fréquence du mot *leadership*. Puisque ce corpus contient essentiellement des débats politiques, le mot *leadership*, utilisé comme caractéristique décrivant une personne, y revient souvent.

la mémoire de traduction d'une entreprise (ici, la Chambre des communes du Canada) en vue de régler des problèmes lexicographiques particuliers. En travaillant régulièrement avec le Hansard, on remarque combien, dans l'ensemble, ces textes sont bien traduits. Les traducteurs à la Chambre des Communes, qui pourtant traduisent des milliers de mots chaque jour, arrivent à le faire de façon idiomatique, voire originale. Le bitexte contient des trouvailles de traduction qu'un bi-concordancier permet d'exploiter facilement.

Le lexicographe doit néanmoins s'en servir avec discernement. En effet, les résultats de TransSearch sont parfois plus utiles à la fin de l'étape de rédaction d'une entrée lexicale (pour confirmer des équivalents, par exemple) ou parfois plus pertinents à la phase dépouillement pour donner des idées d'équivalents. Chaque cas est différent. Par ailleurs, certains mots ou combinaisons de mots se prêtent beaucoup mieux que d'autres à la recherche avec TransSearch. En effet, plus l'unité lexicale recherchée est complexe et figée (expressions figées, composés, appellations officielles et certains types de collocations), plus il est utile car, en règle générale, l'équivalent se trouve là, noir sur blanc. À l'inverse, moins l'unité lexicale est complexe ou figée (mots simples utilisés en combinaisons libres), moins le bi-concordancier est utile, sauf dans la mesure où il peut fournir des équivalents de l'unité lexicale dans des combinaisons libres.

Malgré les quelques réserves mentionnées ci-dessus, il est devenu évident que l'utilisation des bi-concordanciers pour exploiter des bitextes est un élément essentiel de la lexicographie bilingue; les bitextes et les bi-concordanciers devraient faire partie de l'arsenal du lexicographe, au même titre que les corpus unilingues et les concordanciers. Le poste de travail du lexicographe bilingue de l'an 2000 ne saurait s'en passer.

Chapitre 3 : *LES COLLOCATIONS ET LES OUTILS INFORMATIQUES*

Introduction

Nous avons déjà mentionné, au chapitre précédent, l'utilité d'une bi-concordance dans le repérage des collocations. Cependant, l'exploitation d'un bitexte avec un bi-concordancier ne permet pas de repérer systématiquement « les mots qui apparaissent habituellement ensemble » ainsi que leurs traductions. Dans le présent chapitre, nous allons examiner d'autres outils informatiques qui permettent, dans une certaine mesure, d'extraire automatiquement des collocations. Avant de passer aux outils informatiques comme tels, nous tenterons d'abord de préciser ce qu'on entend au juste par *collocation* et d'en présenter quelques typologies. Ensuite, nous expliquerons pourquoi il est primordial de les inclure dans tout dictionnaire bilingue.

3.1 Collocations

3.1.1 Désignation, définition et nature des collocations

Comme beaucoup de termes en linguistique et en lexicographie, le mot *collocation* ne fait l'unanimité ni quant à sa désignation ni quant à sa définition.

Le terme *collocation* a été proposé par J.R. Firth (1951) lorsqu'il affirme que « words shall be known by the company they keep ». Bien que le terme *collocation* soit ancré dans l'usage en

anglais, les linguistes français, eux, ont été aux prises avec une série quasi-synonymique de termes pour exprimer l'idée proposée par Firth. E. Lipshitz (1981), par exemple, utilise tour à tour *phraséologie* et *phraséologismes*[1] pour parler de groupements qui ne sont pas créés au fur et à mesure des besoins, mais qui sont reproduits intégralement par l'usager, car ils sont formés d'avance. Peut-être certains francophones voulaient-ils éviter le terme *collocation*, de peur de calquer le terme anglais. Même si Mounin a inclus le terme *collocation* dans son dictionnaire de 1974[2], il a mis le lecteur en garde : ce terme est « surtout employé par les linguistes anglais »[3]. Depuis quelques années, toutefois, le terme *collocation* semble s'être infiltré dans les textes des linguistes et lexicographes français. Le numéro spécial de la revue Meta[4] intitulé « Hommage à Bernard Quemada – Termes et textes », par exemple, contient deux articles[5] traitant précisément de la *collocation*[6]. C'est pourquoi, dans le cadre de cette thèse, le mot *collocation* sera utilisé en français pour désigner ce que les linguistes anglais appellent *collocation*.

---

1 Il utilise aussi le terme *groupement phraséologique* pour nommer les structures du type SV + SN.

2 Collocation : « dénote l'association habituelle d'une unité lexicale avec d'autres unités ».

3 Fait notable, le mot *collocation* n'apparaît pas dans les dictionnaires de Martinet (1973) et Phelizon (1976).

4 Vol. 39, n° 4.

5 « Tournoi pour l'accommodement des dictionnaires de collocations » (Jean-Luc Descamps 1994) et « Collocations et langue de spécialité » (André Clas 1994).

6 Ceci ne veut pas dire que tous les langagiers francophones sont d'accord. À preuve, après la communication sur la collocation que j'ai présentée lors du congrès annuel de l'Association canadienne de traductologie en mai 1996, quelques membres de l'auditoire m'ont reproché l'utilisation du terme *collocation*. Selon eux, le terme français *cooccurrent* serait le mot juste.

Quoique Firth ait été le premier à utiliser le terme *collocation*, il ne l'a pas défini précisément. Si Mario Pei et Frank Gaynor (1954) ont défini ce terme dans leur dictionnaire, la définition qu'ils proposaient à l'époque, « arrangement of words in a sentence in order to properly convey the intended meaning », ne correspond pas à la définition actuelle du terme. Inspirés par Firth, les linguistes contemporains définissent maintenant la *collocation* en termes de « habitual co-occurrence of individual lexical items » (Crystal 1991), de « 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) ou encore de « co-occurrence of two or more lexical items as realizations of structural elements within a given syntactic pattern » (Anthony Cowie 1978:132).

Ces « recurrent word combinations » (Morton Benson *et al.* 1986:vii), qui sont constitués de mots qui s'attirent l'un l'autre[7], sont un phénomène de langue. De par la nature même des collocations, les locuteurs les récupèrent de leur mémoire en bloc. Selon Sinclair (1991:110), « the language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices ». Dès lors qu'un locuteur veut rendre l'idée de « commencer », par exemple, il doit d'abord préciser le sujet qui fait l'action de *commencer* ou l'objet de cette action et ensuite choisir le verbe qui exprime effectivement l'idée de commencer dans ce contexte précis. Ainsi,

---

[7] Margaret Cop (1988) compare l'attraction entre les mots à des particules chargées électriquement. Les particules qui n'ont aucune charge électrique sont comme les composants d'une combinaison libre (*manger une pomme*) : ils ne s'attirent ni se repoussent. Les composants d'une collocation (*commettre un crime*) agissent comme des particules de charge opposée : ils s'attirent l'un l'autre. Finalement, les particules de même charge rappellent les cas où les auteurs désautomatisent les collocations afin de créer un effet chez le lecteur (*commettre un article*) : ces mots se repoussent.

on *ouvre* un procès, l'orage *éclate*, des pourparlers sont *entamés* et on *entonne* une chanson. Ces unités, que Maurice Pergnier (1980:307) appelle *unités idiomatiques* par rapport à *unités structurales*, sont caractérisées par deux traits, « construction et propriété d'une langue donnée » (Pergnier 1980:312), ce qui fait donc que « l'idiosyncrasie de la collocation ne se révèle définitivement que dans l'optique d'une autre langue qui combine, pour exprimer le même fait, des mots différents » (Franz Hausmann 1990:1013).

3.1.2 Types de collocations

Les chercheurs caractérisent les collocations selon divers critères. D.J. Allerton (1984), qui n'utilise pas le mot collocation comme tel mais parle plutôt de *cooccurrences*, les classe par niveau de restrictions : syntaxiques, sémantiques et locutionnelles. La première catégorie comprend exclusivement les combinaisons tributaires de la syntaxe[8]. Allerton poursuit en montrant que l'acceptabilité de certaines combinaisons est régie au niveau de la sémantique puisque les traits sémantiques d'un mot sont en conflit avec ceux d'un autre[9]. Le troisième type de restrictions, qu'il a appelées *locutional* en s'inspirant des « locutions toutes faites » de Ferdinand de Saussure (1964:172), sont arbitraires et imposées par la langue. Il inclut dans cette catégorie des

---

[8] Par exemple, les règles de la syntaxe forcent le locuteur à utiliser, dans une phrase donnée, un pronom relatif plutôt qu'un autre (*la pomme que j'ai mangée* versus *la pomme dont j'ai mangé*).

[9] Un des exemples expliqués par Allerton est la phrase suivante : *the old spinster had often married.* Cette phrase est tout à fait acceptable sur le plan de la syntaxe, mais les traits définitoires de *spinster* (une femme qui ne s'est jamais mariée) et *to marry* entrent en contradiction l'un avec l'autre.

combinaisons du type Verb+Prép[10] (*to rely on*), NomC+Prép (*faith in*), des verbes très fréquents qui ne prennent sens que lorsqu'ils sont combinés avec un substantif (*to do, to get, to give, to have, to make, to put, to take*), des verbes qui, lorsqu'ils sont associés à un nom, peuvent être remplacés par la forme verbale du nom en question (*to exert influence = to influence*). C'est surtout les deux dernières catégories, qui sont régies par des restrictions sémantiques et locutionnelles, qu'on nomme généralement *collocation*.

Cowie (1978) propose la typologie *open collocation* et *restricted collocation*. Dans la première catégorie, il inclut des combinaisons de mots qui, sur le plan de la sémantique, sont tellement généraux qu'ils peuvent être combinés de façon quasi-illimitée. Le verbe *to run*, par exemple, se combine avec presque tout : *machine, business, horse, program, etc.* Par *collocations restreintes*, Cowie entend des combinaisons de mots dans lesquelles le sens du mot A limite les possibilités pour l'élément B (par exemple, *to explode a claim* ou *to kill a cigarette*). Ces derniers sont beaucoup plus intéressants sur le plan de la lexicographie.

À son tour, Sinclair (1991:115) propose une autre façon de classifier les collocations. Il définit *upward* et *downward collocations* en fonction de la fréquence des composants de la collocation[11].

---

[10] Les abréviations suivantes seront adoptées dans le présent travail : *NomC* (substantif), *Verb* (verbe), *AdjQ* (adjectif qualificatif), *Adve* (adverbe) et *Prép* (préposition).

[11] On appelle le mot principal de la collocation la *base*, ou *node* en anglais, et le(s) mot(s) qui l'accompagne(nt), *collocatif*, ou *collocator* ou *collocate* en anglais. Pour Hausmann (1990:1010), « la base de la collocation [est] le partenaire caractérisé [...] et le collocatif, le partenaire caractérisant qui ne reçoit son identité sémantique que par la collocation ». Certains, comme Sinclair (1991), appellent *node* le mot qui fait l'objet de la recherche et *collocate* le mot qui l'accompagne, nonobstant lequel des deux est sémantiquement fort.

Prenons, par exemple la collocation anglaise *to give an audience*, et supposons que les mots *to give* et *audience* reviennent 500 fois et 30 fois respectivement dans le corpus. Pour Sinclair, la relation collocationnelle entre ces deux composants change en fonction du mot à l'étude. Si le lexicographe s'intéresse au verbe *to give* et trouve que ce mot est fortement associé à *audience*, cette combinaison sera une *downward collocation* puisque le mot *audience* est moins fréquent que le mot à l'étude, soit *to give*. En revanche, si le lexicographe s'était intéressé au mot *audience* plutôt qu'à *to give*, la collocation aurait été du type *upward* puisque le verbe *to give* est plus fréquent que le substantif *audience*.

Comme nous l'avons vu, les chercheurs, en général, caractérisent les collocations selon leurs propres besoins. Ainsi, dans l'optique lexicographique, nous adopterons la typologie très simple proposée par Benson *et al.* (1986). Ils divisent les collocations en deux grands groupes, les **collocations grammaticales** et les **collocations lexicales**. La collocation grammaticale, que Hausmann (1990:1013) appelle *construction*, est constituée d'un mot dominant suivi d'une unité subordonnée (souvent une préposition ou une structure grammaticale, comme un infinitif ou une proposition). Les collocations grammaticales se présentent habituellement sous les formes suivantes (NomC = nom commun et AdjQ = adjectif qualificatif) :

Exemples de collocations grammaticales
Verb+Prép *to abstain from, s'abstenir de* NomC+Prép *anger at, sentiment envers* AdjQ+Prép *absent from, absent de*

Tableau 26– *Exemples de collocations lexicales*

Contrairement à la collocation grammaticale, la collocation lexicale, elle, est généralement formée de deux composantes lexicales d'importance plus ou moins égale. Typiquement, les collocations lexicales sont formées de noms (NomC), d'adjectifs(AdjQ), de verbes(Verb) et d'adverbes(Adve). En voici quelques exemples :

Exemples de collocations lexicales
NomC + Verb *bombs explode, l'orage éclate* AdjQ + NomC *a confirmed bachelor, un célibataire endurci* Verb + NomC *reject an appeal, interjeter appel*

Tableau 27 – *Exemples de collocations lexicales*

Dans cette thèse, nous nous concentrerons sur les collocations lexicales puisque les collocations grammaticales se retrouvent souvent dans les grammaires et les dictionnaires et se repèrent assez aisément dans une concordance. Les collocations lexicales, par contre, sont plus négligées dans les ouvrages de référence et plus difficiles à repérer dans une concordance. Étant donné que seules les collocations lexicales seront à l'étude ici, nous utiliserons désormais le terme « collocation » pour désigner uniquement la *collocation lexicale*.

3.1.3 Identification des collocations

Si les collocations sont difficiles à repérer, c'est qu'elles se situent dans une zone floue, ce que Thierry Fontenelle (1994:45) appelle le *fuzzy area* entre la combinaison libre[12], le composé[13] et l'expression figée[14], dont les frontières sont loin d'être étanches. Nous adopterons les critères de Liang (1991) pour comparer la collocation à l'expression figée, puis à la combinaison libre. Ensuite, nous la comparerons aux composés.

---

[12] On appelle *combinaisons libres* les syntagmes dans lesquels les mots se combinent librement pour former un nombre infini de combinaisons, sans restrictions syntaxiques ou sémantiques. Leur sens est prévisible : c'est la somme du sens des mots dont il est composé (ex: *I am carrying my suitcase* = je + transporte + une valise). Ces mots se combinent conformément aux règles générales de la syntaxe.

[13] Crystal (1991) définit un composé de la façon suivante : « compound : linguistic unit which is composed of elements that function independently in other circumstances ».

[14] Crystal (1991) définit expression figée comme suit : « idiom : sequence of words which is semantically and often syntactically restricted, so that they function as a single unit (...) meanings of the individual words cannot be summed to produce the meaning of the 'idiomatic' expression as a whole ».

Selon Liang, la collocation se démarque de l'expression figée sur trois plans : l'autonomie des composants, leur inaltérabilité sémantique et la substitution possible du collocatif.

Liang considère que les composants d'une collocation sont autonomes parce que, à l'inverse de leur rôle dans une expression figée, chacun conserve sa fonction grammaticale. De plus, on peut les manipuler sur le plan syntaxique (*rédiger une dissertation* et *la dissertation a été rédigée*)[15] ou encore insérer d'autres mots entre les composants (*rédiger d'ici le mois prochain la meilleure des dissertations*). Certes, toutes les collocations ne font pas preuve d'autonomie au même degré, mais il n'en reste pas moins qu'elles se démarquent des expressions figées à ce chapitre. Selon Liang, ce critère est le plus important des critères présentés ici.

Une autre caractéristique des composants d'une collocation est qu'ils conservent leur sens, propre ou figuré, tandis qu'une expression figée adopte globalement un sens figuré ou métaphorique. Ainsi, la collocation présente une certaine transparence. Cela dit, ce critère ne suffit cependant pas pour différencier les collocations des expressions figées, certaines collocations, comme *to curry favour*, pouvant aussi être opaques.

Finalement, Liang propose que les collocations se démarquent des expressions figées par le fait qu'il est assez souvent possible de substituer un collocatif par un autre sans changer le sens de la

---

[15] Sinclair (1991:111) cite aussi cette caractéristique des collocations et donne l'exemple suivant : « set x on fire » et « set fire to x », ou encore « x is not in his nature » et « it is not in his nature to x ».

collocation, par exemple, *jeter/établir/poser/asseoir les bases de quelque chose* (Liang 1991:153), phénomène que Hausmann (1990:1010) appelle le *non-figement* d'une collocation. Comme c'était le cas pour la caractéristique précédente, cependant, certaines expressions figées permettent aussi ce genre de substitution, notamment *ne demander/ne chercher/ne rêver que plaies et bosses* (Liang 1991:153).

Nous avons montré qu'il est difficile de faire la distinction entre une collocation et une expression figée. Néanmoins, le problème est encore plus complexe quand il s'agit de différencier les collocations des combinaisons libres, car elles prennent parfois la même forme. Le syntagme *to deliver a speech*, par exemple, peut avoir deux sens. En combinaison libre, il veut dire *prendre un discours et aller le porter à quelqu'un*. Toutefois, ce même syntagme peut aussi vouloir dire *prononcer un discours*, lorsque nous avons affaire à une collocation. Liang (1991:153) suggère que ce sont des contraintes lexicales crées par l'usage qui opèrent sur les collocations[16] et qui les distinguent des combinaisons libres. Même s'il est possible de remplacer un mot par un autre à l'intérieur d'une collocation, les possibilités ne sont pas illimitées, comme c'est le cas dans les combinaisons libres[17]. Ainsi, les combinaisons qui sont des collocations auront plus tendance que les combinaisons libres à se retrouver dans un corpus, ce qui expliquerait pourquoi la notion de fréquence est si importante dans le repérage automatique des collocations.

---

[16] Ce que Hausmann (1990:1010) appelle *affinité*.

[17] Aisendstadt (1979) partage cet avis.

Le composé et la collocation adoptent souvent la même forme sur le plan grammatical (*NomC de NomC*, *NomC à NomC*, *NomC AdjQ*, *AdjQ NomC* et *NomC à Verbe*, par exemple). Si pareilles combinaisons se repèrent facilement dans un texte étiqueté, il est ensuite très difficile de différencier la collocation du composé. Pour y arriver, de nombreux terminologues utilisent comme critère principal le fait que, contrairement à la collocation, le composé forme un tout sémantique ayant sa propre définition et ses propres caractéristiques. D'autres, comme Roberts (1994/1995), préfèrent différencier les collocations des composés en se servant de la catégorie grammaticale. Ainsi, si la combinaison de mots à l'étude fonctionne comme un substantif, on a affaire à un composé. Si chaque mot de la combinaison conserve plutôt sa propre catégorie grammaticale, cette combinaison est une collocation. Que l'on utilise une façon ou l'autre de séparer les collocations des combinaisons, il n'en reste pas moins que, pour l'instant, la tâche est très difficile.

3.1.4 Importance des collocations dans les dictionnaires bilingues

Étant donné la nature complexe des collocations, on s'attendrait à ce qu'il existe de nombreux dictionnaires de collocations. Néanmoins, pour la paire de langues qui nous intéresse, soit l'anglais et le français, un seul dictionnaire par langue n'a été publié, soit le BBI (Benson *et al.* 1986) en anglais et le Lacroix (1956) en français (bien que ce dernier soit très vieux). C'est pouquoi Geoff Barnbrook (1996:135) suggère qu'une entrée de dictionnaire devrait contenir, en plus d'une définition, des renseignements sur l'environnement habituel de ce mot, surtout lorsque cet environnement est particulier ou non prévisible. Cette recommandation s'applique davantage

encore au dictionnaire général bilingue, car les locuteurs ont encore plus de difficultés à maîtriser les collocations d'une langue seconde.

On consulte généralement un dictionnaire bilingue pour deux raisons : pour décoder un texte en langue de départ ou pour en produire un en langue d'arrivée. Placé devant une collocation inconnue, un locuteur natif, une personne bilingue et même un apprenant arrivent généralement assez facilement à comprendre son sens, car ses composants conservent leur sens individuel. Mais il existe des collocations du type *to curry favour*, celles que Cowie appelle *restricted collocations*, qui s'appréhendent plus difficilement, même par les locuteurs natifs. Pour en connaître le sens, on peut toujours chercher cette collocation dans un dictionnaire unilingue anglais, mais il est parfois plus utile, surtout pour un apprenant, d'en connaître l'équivalent dans sa langue maternelle[18]. Cependant, à l'exception d'un nombre limité de collocations, celles-ci sont transparentes, même pour les apprenants.

En revanche, les collocations posent problème dans l'encodage des textes, surtout pour un apprenant, car il ne peut savoir à l'avance quelles combinaisons contenant un mot donné sont acceptables et lesquelles sont à proscrire. En anglais, par exemple, les adjectifs *strong* et *powerful* sont des quasi-synonymes; cependant, le locuteur anglais ne dira jamais spontanément *powerful tea*. *Powerful tea* serait certes compris par tout locuteur anglais, mais cette combinaison serait

---

[18] Pour un francophone, par exemple, la définition du Random House Webster's College Dictionary, *to seek to advance oneself through flattery or fawning*, peut être difficile à comprendre. Par contre, il saura exactement ce que veut dire *to curry favor* lorsqu'il lira l'équivalent proposé dans le Robert-Collins Senior, *chercher à gagner la faveur de quelqu'un*.

perçue comme n'étant pas idiomatique. D'où l'importance de la place que les collocations doivent tenir dans tout dictionnaire, surtout un dictionnaire bilingue. Par exemple[19], pour rendre en anglais l'idée qu'*un orage a éclaté*, un locuteur anglais ne dirait jamais *the storm has exploded* comme le voudrait la traduction littérale. Mais à l'entrée *éclater*[20] du Robert-Collins Senior, on trouve, pour la collocation *orage+éclater*, la collocation équivalente en anglais, *storm +to break*[21].

Pour les raisons évoquées ci-dessus, et bien d'autres encore, il est important d'inclure, dans un dictionnaire bilingue, autant de collocations que possible. Cependant, la tâche du lexicographe bilingue est suffisamment ardue sans lui imposer le devoir supplémentaire de repérer, à partir d'un corpus, des collocations d'un mot-vedette et de ses équivalents. C'est pour simplifier quelque peu son travail que nous avons pensé à produire un outil qui permettrait d'extraire semi-automatiquement des collocations et leurs traductions.

---

[19] Cet exemple est tiré d'une communication présentée en mai 1996 dans le cadre du congrès annuel de l'Association canadienne de traductologie.

[20] Nous ne discuterons pas, dans le cadre de cette thèse, de l'entrée à laquelle l'information devrait être présentée. Pour une explication plus approfondie, se référer à Cop (1988).

[21] Par exemple, les lexicographes bilingues utilisent les collocations pour désambiguïser les sens de certains mots. Le verbe *éclater*, dans le Robert-Collins Senior, comprend six divisions sémantiques. Deux seulement contiennent une indication de sens, soit les sens c) *retentir* et d) *se manifester*. Pour les deux premières divisions sémantiques, par exemple, ce n'est qu'avec les collocatifs qu'on peut inférer le sens du verbe (collocatifs du sens a : *[mine, bombe], [obus], [veine], [bourgeon], [pneu, chaudière], [verre], [parti, ville, services, structures familiales]*; collocatifs du sens b : *[incendie, épidémie, guerre], [orage, scandale, nouvelle]*).

3.2 Description de Xtract

Il existe plusieurs logiciels du type *analyseur de textes* qui permettent, jusqu'à un certain point, de repérer des collocations dans un corpus unilingue[22]. Mais peu de logiciels sont conçus particulièrement pour ce but. Le plus élaboré à l'heure actuelle est sans doute Xtract, qui fera l'objet de la présente section. Ce logiciel, conçu par Frank Smadja (1993), permet de repérer des collocations de longueur variable, que les mots soient contigus ou non[23]. Il faut noter que Smadja définit le terme *collocation* de façon assez vague. En effet, pour lui, la collocation comprend tout ce que son logiciel récupère, soit des collocations telles que définies à la section 3.1, soit des phrases du type *The Dow Jones industrial rose xxx points* (que Smadja appelle des *phrasal templates*) ainsi que des composés comme *The Dow Jones industrial* (des *rigid noun phrases* pour Smadja).

Xtract fonctionne en trois étapes. Il repère d'abord, à partir d'un corpus étiqueté avec les catégories grammaticales, les paires de mots (bigrammes) statistiquement significatives[24]. Les composants de ces paires peuvent être séparés l'un de l'autre par un maximum de 4 mots. À la prochaine étape, Xtract extrait, à partir des bigrammes déjà trouvés, des unités significatives plus

---

[22] Des concordanciers comme MicroConcord, TACT et PAT, par exemple.

[23] Le CITI nous a permis d'utiliser sa version de Xtract sur une année du Hansard. Ce logiciel est particulièrement difficile à manipuler pour un non-programmeur, et l'utilisateur doit avoir des bonnes bases en Unix. Aussi, nous ne ferons pas état de nos résultats ici, mais nous contenterons de citer ceux de Smadja.

[24] La notion de "statistiquement significative" sera expliquée à la prochaine section.

longues (n-grammes), ce qu'il appelle *phrasal templates* ou les *rigid noun phrases*. En dernier lieu, il élimine des bigrammes trouvés à la première étape toutes les combinaisons dans lesquelles les composants ne conservent pas toujours la même relation syntaxique. Supposons, par exemple, que Xtract ait repéré le couple *price* + *rose*. Les relations syntaxiques qui peuvent exister entre ces mots pourraient être : *sujet* + *verbe* (*the prices rose...*), *verbe* + *complément* (*to price a rose*), ou encore *substantif* + *complément* (*the price of a rose*). Ainsi, la combinaison *price* + *rose*, si elle avait été trouvée, aurait pu être éliminée à cette étape-ci[25].

Seule la première étape, soit l'extraction de bigrammes, sera à l'étude ici[26] parce que c'est surtout en bigrammes que se manifestent les collocations et que ce seront uniquement les bigrammes qui seront traités dans le cadre de cette thèse.

Extraction des bigrammes

L'extraction des bigrammes par Xtract se fait en trois sous-étapes qui seront résumées ici. Xtract segmente d'abord le texte en phrases en se servant de la ponctuation marquant la fin des phrases[27]. Seules les phrases qui contiennent le mot *w* demandé sont retenues (tableau 28).

---

[25] Le scénario que nous avons décrit serait raisonnable si Xtract avait tourné sur un corpus général. Mais puisqu'il ne tournait que sur un corpus traitant des marchés boursiers, la collocation *price* + *rose* a été repérée par Xtract, puisque, dans le contexte des marchés boursiers, il ne peut y avoir aucune ambiguïté au niveau de l'analyse grammaticale.

[26] La description qui suit résume la partie pertinente de l'article de Smadja (1993). Tous les résultats présentés par Smadja et que nous reproduisons ici sont extraits d'un corpus de dix millions de mots traitant des marchés boursiers.

[27] Quelques erreurs se glissent à ce stade à cause justement de la difficulté de repérer ces points (confusion possible avec les abréviations, les acronymes, etc.)

Étape 1.1 – Production de concordances[28]
Entrée : Un corpus étiqueté et un mot *w* au choix de l'usager. **Sortie :** Toutes les phrases contenant le mot *w*.

Tableau 28 – *Xtract : étape 1.1 – production de concordances*

Xtract dresse ensuite une liste de tous les collocatifs w_i du mot *w* et, pour chaque w_i, tient compte de sa position par rapport à *w*, de sa catégorie grammaticale et ainsi que de sa fréquence. À ce stade-ci, les mots vides ou grammaticaux et les mots séparés de *w* par plus de cinq mots sont éliminés (tableau 29).

Étape 1.2 - Compilation et tri[29]
Entrée : La sortie de l'étape 1.1, soit des phrases contenant le mot *w*. **Sortie :** La liste des collocatifs w_i et de leur fréquence avec le mot *w*.

Tableau 29 – *Xtract : étape 1.2 – compilation et tri*

[28] Traduction de *Producing Concordances*, le terme employé par Smadja.

[29] Traduction de *Compile and Sort*, le terme employé par Smadja.

Le tableau 30 montre les résultats de l'analyse d'une seule phrase, présentés dans l'article de Smadja (1993:152).

Collocatifs de *takeover* dans la phrase *The pill would make a takeover attempt more expensive by allowing the retailer's shareholders to...*			
w	***w_i***	**Distance[30]**	**Catégorie grammaticale**
takeover	pill	-4	N
takeover	make	-2	V
takeover	attempt	+1	N
takeover	expensive	+3	Adj
takeover	allowing	+5	V

Tableau 30 — *Xtract : collocatifs de takeover*

Dans cette phrase, par exemple, les w_i significatifs sont *pill* (distance +4), *make* (distance +2), *attempt* (distance -1), *expensive* (distance -3) et *allowing* (distance -5). Xtract analyse de cette façon toutes les phrases contenant le mot *takeover*.

Ensuite, Xtract estime pour chaque mot *w* la fréquence moyenne de tous ses collocatifs w_i ainsi que l'écart type σ pour ces fréquences. Puis, il calcule la *force* du lien (k_i) pour chaque paire *w*-w_i

[30] Dans ce tableau, Smadja (1993:154) utilise une distance *positive* pour un mot placé avant *w* et distance *négative* pour un *w* placé après. Néanmoins, partout ailleurs, il utilise la notation inverse. Pour éviter toute confusion dans cette thèse, une valeur positive voudra dire que w_i vient après *w* et une valeur négative, qu'il vient avant.

en soustrayant la fréquence moyenne de la fréquence à une distance donnée et en divisant ensuite par l'écart-type[31] (tableau 31).

Étape 1.3 - Analyse[32]
Entrée : La sortie de l'étape 1.2, soit la liste des w_i avec, pour chacun, sa fréquence en fonction de sa position relativement à w. **Sortie :** Des paires de mots significatives ainsi que des renseignements sur leur fréquence et leur position relative par rapport au mot demandé.

Tableau 31 – *Xtract : étape 1.3 – analyse*

Le tableau 32 présente les résultats calculés par Xtract pour le mot *takeover* en combinaison avec un adjectif.

---

[31] L'équation utilisée par Smadja est la suivante :

$$\bar{k}_i = \frac{freq_i - f}{\sigma}$$

où k_i est la force du lien entre w et w_i pour une distance d; $fréq_i$ est la fréquence de w_i à cette distance d; f est la fréquence moyenne de tous les w_i à toutes les distances d_i; σ est l'écart type des fréquences f relativement à la moyenne.

[32] Traduction de *Analyze*, le terme employé par Smadja

Sortie de l'étape 1.3 - Combinaisons takeover+adjectif (extrait)												
w	w_i	*Fréq*	p_{-5}	p_{-4}	p_{-3}	p_{-2}	p_{-1}	p_1	p_2	p_3	p_4	p_5
takeover	possible	178	0	13	4	23	138	0	0	0	0	0
takeover	corporate	93	2	2	2	1	63	3	2	9	4	5
takeover	unsolicited	83	5	30	5	0	42	0	0	1	0	0
takeover	several	81	2	6	6	6	45	0	0	12	0	4
takeover	recent	76	5	4	6	5	17	0	0	36	2	1

Tableau 32 – *Xtract : combinaisons takeover + adjectif*

La prochaine phase consiste à évaluer la distribution moyenne (p_i) de w_i autour de *w* pour calculer la variance U_i[33] autour de p_i (échelle de 1 à 100). Plus U_i est petit, plus l'histogramme représentant la distribution du mot w_i relativement à *w* est plat, et donc moins la position de w_i autour de *w* est fixe. À l'inverse, plus U_i est grand, plus w_i a des chances de se retrouver dans un ou plusieurs endroits précis autour de *w* et donc de former des collocations significatives[34].

Voici deux histogrammes fictifs (tableau 33) montrant la distribution d'un mot w_i autour d'un mot *w*. Le premier histogramme est pratiquement plat et n'est que de peu d'intérêt sur le plan collocationnel. Le second, cependant, comporte deux colonnes qui se démarquent clairement des

---

[33] La variance U_i est calculée comme suit :

$$U_i = \frac{\sum_{j=1}^{10} (p_i^j - \bar{p}_i)^2}{10}$$

[34] Smadja utilise ici *collocation significative* de la même façon qu'il a utilisé *significatif* avec *paires*, *unités* et *mots*, c'est-à-dire qui se démarquent des autres collocations (paires, unités ou mots) strictement au plan statistique.

autres, soit celle à la position *p-3* (*p* étant la position de *w*) et *p-2*. Ce deuxième histogramme représente assurément une collocation puisque w_i n'apparaît que dans certaines positions fixes par rapport à *w*.

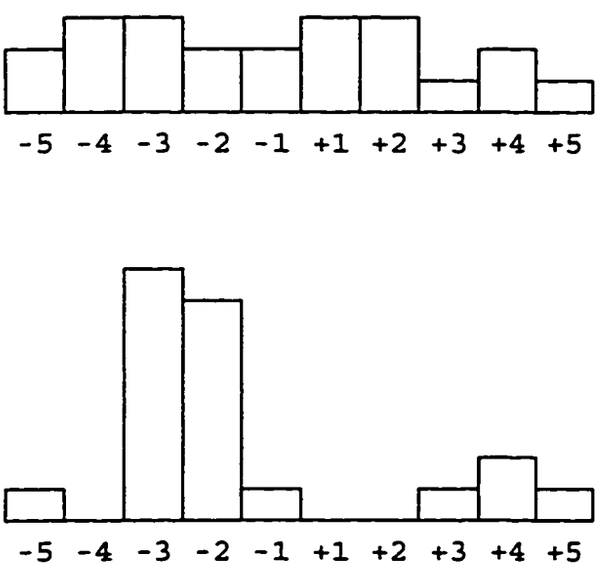

Tableau 33 – *Xtract : histogramme – distribution d'un mot autour d'un autre*

Xtract calcule, pour les positions allant de -5 à +5 autour du mot-vedette, la différence entre le nombre d'occurrences du collocatif pour cette position et la distribution moyenne du collocatif et met cette différence au carré pour neutraliser les résultats négatifs $(p_i^j - p_i)^2$. Le logiciel calcule la moyenne des 10 positions autour du mot-vedette et utilise ensuite ces statistiques pour éliminer certaines paires de mots qui seraient moins significatives. L'utilisateur peut préciser le seuil minimum pour la force du lien k_o, et seules les paires qui ont une valeur supérieure à ce seuil sont retenues.

La possibilité qu'une paire $w+w_i$ forme une collocation significative augmente avec la valeur de U_i. Ainsi, on précise une valeur limite en-deça de laquelle les paires sont rejetées. Ce filtre assure qu'un sommet au moins apparaît dans l'histogramme, ce que Smadja appelle *pattern of coappearance*, c'est-à-dire une ou des positions autour de w privilégiée par w_i.

Finalement, Xtract récupère les pointes de l'histogramme dont la valeur k est supérieure à une valeur k_1 pré-déterminée.

Pour Smadja, dès lors, la relation lexicale s'exprime en tuple, soit (w_i, *distance, force, distribution*,w_j). Voici un exemple de résultats fournis par Xtract à la sortie de cette étape-ci (tableau 34).

Sortie – *takeover*				
w_i	w_j	Distance (*d*)	Force (*k*)	Distribution (*U*)
hostile	takeovers	1	13	96
hostile	takeover	1	13	90
corporate	takeovers	1	8	90
possible	takeover	1	6	73
hostile	takeovers	2	2	70

Tableau 34 – *Xtract : sortie – takeover*

Il s'agit bel et bien de collocations extraites automatiquement d'un corpus monolingue.

3.3 Description de Champollion

Après avoir créé le logiciel Xtract, Smadja (1996) a mis au point (en collaboration avec Kathleen McKeown et Vasileios Hatzivassiloglou), un autre programme, *Champollion*, qui permet d'extraire automatiquement la traduction d'une collocation donnée en entrée. Plus précisément, l'utilisateur donne un mot ou une collocation en entrée, et Champollion trouve le mot ou la combinaison de mots qui est le plus susceptible d'en être la traduction[35].

Balayant un bitexte apparié au niveau des phrases[36], Champollion, tout comme Xtract d'ailleurs, se sert exclusivement de statistiques pour repérer la traduction d'une collocation. Bien que les auteurs n'aient utilisé en entrée que des collocations produites par Xtract, Champollion peut en fait traiter n'importe quelle collocation, qu'elle ait été repérée par Xtract ou pas.

Ce qui suit est une description très sommaire de l'algorithme en question. Nous utiliserons la collocation[37] *official languages* (Smadja *et al.* 1996:17) pour illustrer le fonctionnement de Champollion.

---

[35] Si Champollion prend des mots simples en entrée, c'est qu'ils peuvent être traduits par une collocation en L2 (comme par exemple *attaquer –> to launch an attack*) ou, à l'inverse, qu'une collocation en L1 pourrait être traduite par un mot unique en L2.

[36] Afin d'éviter des erreurs qui seraient directement attribuables à des problèmes d'alignement du bitexte, Smadja et ses collègues n'ont utilisé que des phrases pour lesquelles l'appariement était 1 à 1.

[37] Le mot *collocation* est utilisé ici au sens où Smadja l'entend, c'est-à-dire que ce terme comprend ce que d'autres appellent des *noms composés*.

Dans un premier temps, Champollion prend en entrée une collocation anglaise et repère, dans la partie française du bitexte, tous les mots qui lui sont fortement associés. De ces mots, il ne conserve que ceux dont la valeur du coefficient de Dice (CD), est au-dessus du seuil précisé par l'utilisateur dans les paramètres initiaux. Cette liste de mots sera désignée *liste A* dans le passage suivant.

Champollion suppose que la collocation équivalente à la collocation de départ sera nécessairement composée de mots qui figurent sur la liste A. C'est pourquoi il produit des paires de mots en combinant entre eux tous les mots de la liste A et calcule ensuite le CD pour chaque paire. Ici encore, le logiciel dresse la liste des paires les plus significatives[38] (liste B) et retient la paire dont le CD est le plus élevé.

Champollion répète cette dernière étape et passe des paires significatives aux triplets significatifs, puis des triplets significatifs aux groupements de quatre mots, et ainsi de suite, retenant chaque fois le groupement dont le CD est le plus élevé et faisant passer à la prochaine itération les groupements dont le CD dépassait le seuil établi. Champollion n'arrête que lorsqu'il ne trouve plus de combinaisons dont le CD dépasse le seuil.

Revenons à l'exemple de Smadja, *official languages*, Champollion a d'abord repéré ces mots dans le texte source et n'a conservé que les phrases dans lesquelles ces mots apparaissaient ensemble.

---

[38] Smadja *et al.* entendent par significative celles dont le CD dépasse le seuil établi au départ par l'utilisateur.

Les mots *official* et *languages* sont revenus dans le corpus 492 et 266 fois respectivement, dont 167 fois dans la même phrase. Champollion a calculé le CD pour chaque mot compris dans les phrases françaises qui avaient été alignées avec ces phrases anglaises. Onze mots seulement avaient un CD supérieur au seuil établi (liste A). Les auteurs ne précisent pas quels mots figurent sur cette liste, hormis *officielles*, car le CD de ce mot était le plus élevé de la liste A. Ce mot, ainsi que tous les mots composant la liste A, ont été traités à l'étape suivante. Champollion avait constitué des paires de mots en combinant ensemble les onze mots formant la liste A. Des 35 paires dont le CD dépassait le seuil (liste B), la paire *officielles langues* avait la valeur la plus élevée. Cette paire a donc été retenue, et toutes les paires figurant sur la liste B ont été traitées de la même façon pour obtenir des triplets. Le triplet le plus significatif a été retenu, et tous les triplets ont été traités de la même façon. Champollion a répété cette étape jusqu'à ce qu'aucune combinaison de mots (qui apparaissaient à l'origine sur la liste A) n'ait un CD supérieur au seuil établi. Le tableau 35 montre la combinaison qui a été retenue à chaque itération.

Traductions possibles		
Combinaisons	**CD**	**Nombre de combinaisons**
officielles	0,91	11
officielles langues	0,95	35
honneur officielles langues	0,45	61
déposer honneur officielles langues	0,36	71
déposer pétitions honneur officielles langues	0,34	56
déposer lewis pétitions honneur officielles langues	0,32	28
doug déposer lewis pétitions honneur officielles langues	0,32	8
suivantes doug déposer lewis pétitions honneur officielles langues	0,20	1

Tableau 35 – *Champollion* : *résultats*

Champollion choisit maintenant la combinaison dont le CD est le plus haut, ici *officielles langues*, vérifie dans le corpus de quelle façon cette combinaison significative se présente afin de rétablir l'ordre habituel, si nécessaire. Fait à noter, Champollion fera la distinction entre une combinaison *flexible*, si l'ordre des mots qui la composent est variable, et *rigide* lorsque l'ordre est fixe.

Cette approche, aussi élaborée soit-elle, pose certains problèmes dès le départ. Pour les auteurs, une collocation est non ambiguë en langue source et possède une traduction unique en L2, « in a clear majority of cases » (Smadja *et al.* 1996:7).

Toutefois, tous ceux qui se sont le moindrement attardés aux collocations savent que ni l'une ni l'autre de ces prémisses n'est vraie. Par exemple, la collocation *to receive a citation* a deux sens tout à fait différents en anglais selon le sens du mot *citation* comme tel. En effet, elle peut vouloir dire que la personne a reçu une mention honorable ou encore un ordre pour apparaître en cour.

Quant à la traduction unique d'une collocation, rien n'est plus loin de la vérité. Par exemple, la collocation anglaise *to take action* peut être traduite tour à tour par *agir*, *passer à l'action*, et *prendre des mesures*, pour ne citer que trois synonymes. Et puisque Champollion ne propose que la combinaison dont le CD est le plus élevé, un seul de ces équivalents sera retenu.

De plus, même si le bitexte utilisé est anglais-français, les auteurs n'ont utilisé leur modèle que sur des collocations anglaises. Il serait intéressant de savoir dans quelle mesure le modèle est réversible.

Chapitre 4 : *LA CRÉATION D'UN EXTRACTEUR DE COLLOCATIONS BILINGUES*

Introduction

Puisqu'il n'y a, à notre connaissance, qu'un seul logiciel qui soit destiné à l'extraction automatique des collocations dans les bitextes et que ce logiciel présente certaines limites, nous avons décidé d'essayer d'en créer un autre. À cette fin, nous avons utilisé trois programmes différents, qui seront présentés dans la première section de ce chapitre. La façon dont ils ont été jumelés pour en arriver à un outil intéressant en lexicographie bilingue sera expliquée en détail dans la deuxième section. Enfin, dans la dernière partie du chapitre, nous montrerons de quelle façon nous tentons de réunir en un programme facile à utiliser, toutes les étapes que nous effectuons maintenant manuellement.

Le bitexte dont nous nous sommes servis comme corpus est le *Hansard* apparié selon la méthode à deux temps expliquée à la section 1.2.3 : dans un premier temps, l'alignement a été effectué en fonction de la longueur des phrases et, dans un deuxième temps, en fonction des mots apparentés.

Le bitexte ainsi obtenu a ensuite été étiqueté avec les catégories grammaticales, et seuls les noms communs (*NomC*), les verbes (*Verb*), les adjectifs (*AdjQ*) et les adverbes (*Adve*) ont été conservés dans le bitexte. Cette dernière étape s'avérait nécessaire pour réduire le bruit; puisque les collocations lexicales en français et en anglais ne contiennent que des noms, des verbes, des

adjectifs et des adverbes, il était donc raisonnable d'éliminer les mots des autres catégories grammaticales.

Finalement, une analyse morphologique du bitexte a été effectuée. Ainsi, chaque *mot* a été lemmatisé, c'est-à-dire ramené à sa forme canonique. Cette dernière étape permet de regrouper sous un même lemme toutes les formes que peut prendre un mot, facilitant du coup le repérage de combinaisons en réduisant la dilution de l'information sur plusieurs formes. Les phrases *l'orage éclate*, *l'orage a éclaté*, *l'orage éclatait* et *les orages ont éclaté*, par exemple, apparaîtront toutes, dans un corpus étiqueté et lemmatisé, sous la forme canonique *orage/NomC éclater/Verb*, ce qui fait ressortir l'importance de cette combinaison comparativement à d'autres.

4.1 Description des programmes utilisés

Trois programmes ont servi de base à l'outil que nous avons voulu créer, soit *matrix_consult*, *coloc* et *inf_mut*.

4.1.1 Matrix_consult

Le premier outil informatique, *matrix_consult*[1], s'inspire d'un article de Peter Brown *et al.* (1993) dans lequel les auteurs proposaient cinq modèles statistiques pour la traduction automatique. Le

[1] Le nom de ce programme, baptisé par l'équipe du CITI, ne prend pas la majuscule.

CITI s'est servi du deuxième modèle présenté dans cet article pour élaborer deux dictionnaires bilingues probabilistes, l'un français-anglais et l'autre anglais-français. L'explication qui suit résume l'approche utilisée pour les mots anglais, mais elle est identique à celle utilisée pour les mots français : il suffit de remplacer l'adjectif *anglais* par *français* vice-versa.

L'algorithme de Brown *et al.* calcule la probabilité qu'un mot français soit la traduction d'un mot anglais ou, plus formellement, $Pr(f|e)$, la probabilité d'un mot français f étant donné un mot anglais e. Le programme trouve d'abord toutes les phrases anglaises dans lesquelles e apparaît. Puisque ce modèle traite un bitexte, le fait de repérer les phrases anglaises récupère aussi les phrases françaises avec lesquelles elles sont appariées.

Nous avons donc une série de phrases anglaises contenant e ainsi que les phrases françaises qui ont été jumelées à ces phrases anglaises. Le programme suppose *a priori* que tous les mots français contenus dans ces phrases pourraient être des traductions valables de e.

De plus, Brown *et al.* ont supposé que la probabilité que deux mots alignés soient des traductions l'un de l'autre croît au fur et à mesure que leur position dans leur phrase respective se rapproche l'une de l'autre. Ainsi, si un mot e apparaît au début d'une phrase en anglais, son équivalent français aura aussi tendance à se trouver en début de phrase. Et, à l'inverse, si un mot e se trouve à la fin d'une phrase anglaise, c'est aussi en fin de phrase que se trouvera son équivalent français. Ainsi, pour chaque mot français, le modèle tient compte de la position que celui-ci occupe dans sa phrase comparativement à la position de e dans la phrase anglaise. Le programme calcule, pour

chaque *f*, la ***probabilité*** qu'il soit l'équivalent de *e*, calcul basé d'une part, sur la fréquence de *f* dans les phrases françaises appariées aux phrases contenant *e* et, d'autre part, sur la position de *f* dans la phrase française.

Ce calcul a été effectué pour tous les mots anglais contenus dans le bitexte (plus de 24 millions de mots anglais et 22 millions de mots français), et les résultats ont été stockés sur ordinateur sous forme matricielle. Le tableau 36 montre de façon générique comment la matrice anglaise se présente. Sur chaque ligne de la matrice, on trouve un mot *e* d'une catégorie grammaticale donnée (*cat_gram*) suivi des traductions possibles f_i (associées aussi à une *cat_gram*). Les mots f_i sont présentés en ordre décroissant de probabilité (fl_1 est plus fréquent que fl_2 qui, lui, est plus fréquent que fl_3, et ainsi de suite).

Mot anglais	**Équivalents français possibles**					
e1	$f1_1$	$f1_2$	$f1_3$	$f1_4$	...	$f1_n$
e2	$f2_1$	$f2_2$	$f2_3$	$f2_4$	...	$f2_n$
e3	$f3_1$	$f3_2$	$f3_3$	$f3_4$	...	$f3_n$
...						

Tableau 36 : *Matrice – mots anglais*

La matrice calculée pour les mots français se présente formellement de la même façon, mais les langues sont inversées (tableau 37).

Mot français	Équivalents anglais possibles					
f1	$e1_1$	$e1_2$	$e1_3$	$e1_4$	...	$e1_n$
f2	$e2_1$	$e2_2$	$e2_3$	$e2_4$	...	$e2_n$
f3	$e3_1$	$e3_2$	$e3_3$	$e3_4$	...	$e3_n$
...						

Tableau 37 : *Matrice – mots français*

Le CITI a ensuite élaboré *matrix_consult*, une « moulinette » destinée à consulter cette matrice, ce qui permet de repérer les équivalents probables d'un mot donné. L'utilisateur précise d'abord la matrice qu'il veut interroger (soit la matrice anglais-français ou la matrice français-anglais) et fournit à *matrix_consult*, en entrée, un mot assorti de sa catégorie grammaticale (*NomC*, *Verb*, *AdjQ* ou *Adve*).

Le tableau 38 montre les quarante premiers résultats de la requête pour le mot *mistake/NomC* en anglais.

```
balzac% matrix_consult -wdir all.flx.tag.m1 -rs 'mistake/NomC' -c 0 39 -nc 1
-dnh -dp -p

#================================#=======================================#
# 1391 : mistake/NomC   [NH 2051 ]#    0 : erreur/NomC      0.599409       #
# 1391 : mistake/NomC   [NH 2051 ]#    1 : commettre/Verb   0.211614       #
# 1391 : mistake/NomC   [NH 2051 ]#    2 : tromper/Verb     0.043184       #
# 1391 : mistake/NomC   [NH 2051 ]#    3 : grave/AdjQ       0.014441       #
# 1391 : mistake/NomC   [NH 2051 ]#    4 : faute/NomC       0.008166       #
# 1391 : mistake/NomC   [NH 2051 ]#    5 : méprendre/Verb   0.006905       #
# 1391 : mistake/NomC   [NH 2051 ]#    6 : tort/NomC        0.006874       #
# 1391 : mistake/NomC   [NH 2051 ]#    7 : répéter/Verb     0.006659       #
# 1391 : mistake/NomC   [NH 2051 ]#    8 : corriger/Verb    0.004598       #
```

```
# 1391 : mistake/NomC    [NH 2051 ]#    9 : pas/Adve          0.003491        #
# 1391 : mistake/NomC    [NH 2051 ]#   10 : bévue/NomC        0.003306        #
# 1391 : mistake/NomC    [NH 2051 ]#   11 : monumental/AdjQ   0.003183        #
# 1391 : mistake/NomC    [NH 2051 ]#   12 : éviter/Verb       0.002999        #
# 1391 : mistake/NomC    [NH 2051 ]#   13 : leurrer/Verb      0.002876        #
# 1391 : mistake/NomC    [NH 2051 ]#   14 : avoir/Verb        0.002784        #
# 1391 : mistake/NomC    [NH 2051 ]#   15 : glisser/Verb      0.002722        #
# 1391 : mistake/NomC    [NH 2051 ]#   16 : faux/AdjQ         0.002661        #
# 1391 : mistake/NomC    [NH 2051 ]#   17 : réparer/Verb      0.002168        #
# 1391 : mistake/NomC    [NH 2051 ]#   18 : comporter/Verb    0.002138        #
# 1391 : mistake/NomC    [NH 2051 ]#   19 : ne/Adve           0.001984        #
# 1391 : mistake/NomC    [NH 2051 ]#   20 : doute/NomC        0.001901        #
# 1391 : mistake/NomC    [NH 2051 ]#   21 : mauvais/AdjQ      0.001890        #
# 1391 : mistake/NomC    [NH 2051 ]#   22 : passé/AdjQ        0.001498        #
# 1391 : mistake/NomC    [NH 2051 ]#   23 : contenir/Verb     0.001359        #
# 1391 : mistake/NomC    [NH 2051 ]#   24 : conséquence/NomC  0.001344        #
# 1391 : mistake/NomC    [NH 2051 ]#   25 : premier/AdjQ      0.001321        #
# 1391 : mistake/NomC    [NH 2051 ]#   26 : appeler/Verb      0.001305        #
# 1391 : mistake/NomC    [NH 2051 ]#   27 : reconnaître/Verb  0.001305        #
# 1391 : mistake/NomC    [NH 2051 ]#   28 : route/NomC        0.001301        #
# 1391 : mistake/NomC    [NH 2051 ]#   29 : clair/AdjQ        0.001248        #
# 1391 : mistake/NomC    [NH 2051 ]#   30 : là-dessus/Adve    0.001217        #
# 1391 : mistake/NomC    [NH 2051 ]#   31 : produire/Verb     0.001159        #
# 1391 : mistake/NomC    [NH 2051 ]#   32 : douter/Verb       0.001036        #
# 1391 : mistake/NomC    [NH 2051 ]#   33 : sûr/AdjQ          0.001032        #
# 1391 : mistake/NomC    [NH 2051 ]#   34 : capable/AdjQ      0.000971        #
# 1391 : mistake/NomC    [NH 2051 ]#   35 : responsable/AdjQ  0.000882        #
# 1391 : mistake/NomC    [NH 2051 ]#   36 : laisser/Verb      0.000840        #
# 1391 : mistake/NomC    [NH 2051 ]#   37 : lacune/NomC       0.000817        #
# 1391 : mistake/NomC    [NH 2051 ]#   38 : parfois/Adve      0.000794        #
# 1391 : mistake/NomC    [NH 2051 ]#   39 : illusion/NomC     0.000748        #
# [...]                                                                       #
#================================#=======================================#
```

Tableau 38 : *Ligne de matrice pour le nom 'mistake'*

Une explication de la commande *balzac% matrix_consult -wdir all.flx.tag.ml -rs 'mistake/NomC' -c 0 39 -nc 1 -dnh -dp -p* s'impose. Ici, *matrix_consult* a utilisé le modèle *-wdir* pour repérer, dans la matrice *all.flx.tag.rev.ml*, la ligne à laquelle apparaissaient la forme *mistake/NomC*. Le logiciel n'a sorti que les quarante premiers équivalents (donc les quarante premières colonnes de la matrice), car le paramètre *-c 0 39* indique que seuls les résultats inscrits aux colonnes 0 à 39

doivent être affichés. Le paramètre suivant, soit *-nc 1*, précise que les résultats doivent être présentés sur une colonne, à raison d'un équivalent (et toute l'information s'y rapportant) par ligne. Le paramètre *-dnh* demande au système de donner la fréquence du mot demandé, *-dp*, de numéroter les équivalents, et *-p*, d'afficher les résultats à l'écran.

La partie de gauche du tableau précédent présente l'information sur le mot en langue de départ. Le premier chiffre (1391) indique la ligne de matrice à laquelle se trouve le mot *mistake/NomC*, ce qui veut dire que, en termes de fréquence, le substantif *mistake* se situe au 1391^{e} rang des mots anglais. Ensuite, l'objet de la requête, *mistake/NomC*, est précisé, suivi de sa fréquence (le substantif *mistake* apparaît 2 051 fois dans le corpus).

Dans la partie droite du tableau (séparée de la gauche par une colonne de *#)* sont présentés les renseignements relatifs aux équivalents potentiels. Le premier chiffre indique le rang auquel se trouve l'équivalent (pour des raisons de programmation, l'équivalent le plus probable, soit le premier en tête de liste, est classé au rang 0). Ce rang est suivi de l'équivalent tiré du dictionnaire bilingue probabiliste et de sa catégorie grammaticale. La probabilité comme telle vient ensuite. Ainsi, la probabilité que *erreur/NomC* soit la traduction de *mistake/NomC* est de 0,599409, soit de près de 60 p. 100. Fait à noter, la somme de toutes les probabilités indiquées dans la colonne de droite pour tous les f_i est égale à 1.

La liste des équivalents trouvés dans la matrice est intéressante. Non seulement elle montre des équivalents habituels de *mistake* comme *erreur, faute, tort* et *bévue*, mais elle présente aussi des

verbes comme *tromper* et *méprendre* qui traduiraient très bien, par exemple, *to commit a mistake*. Toutefois, ce qui attire particulièrement l'attention de quiconque s'intéresse aux collocations, ce sont les combinaisons que l'on peut faire avec les différents équivalents qui figurent sur la liste. Par exemple, à *erreur* ou *faute*, on peut associer *commettre*, *grave*, *corriger* et *monumental* qui apparaissent sur la liste, et ces combinaisons sont, à n'en pas douter, des collocations.

Voici maintenant ce que *matrix_consult* a repéré lorsqu'on a lancé une requête pour *erreur/NomC* (tableau 39).

```
balzac% matrix_consult -wdir all.flx.tag.rev.m1 -rs 'erreur/NomC' -c 0 39 -nc 1
-dnh -dp -p

#================================#=======================================#
#  581 : erreur/NomC    [NH 6187 ]#    0 : mistake/NomC        0.286417     #
#  581 : erreur/NomC    [NH 6187 ]#    1 : error/NomC          0.199803     #
#  581 : erreur/NomC    [NH 6187 ]#    2 : make/Verb           0.101009     #
#  581 : erreur/NomC    [NH 6187 ]#    3 : wrong/Adve          0.061885     #
#  581 : erreur/NomC    [NH 6187 ]#    4 : believe/Verb        0.059670     #
#  581 : erreur/NomC    [NH 6187 ]#    5 : think/Verb          0.051550     #
#  581 : erreur/NomC    [NH 6187 ]#    6 : understand/Verb     0.051304     #
#  581 : erreur/NomC    [NH 6187 ]#    7 : be/Verb             0.030635     #
#  581 : erreur/NomC    [NH 6187 ]#    8 : mistake/Verb        0.013057     #
#  581 : erreur/NomC    [NH 6187 ]#    9 : understanding/AdjQ  0.012872     #
#  581 : erreur/NomC    [NH 6187 ]#   10 : wrong/AdjQ          0.008505     #
#  581 : erreur/NomC    [NH 6187 ]#   11 : correct/Verb        0.006967     #
#  581 : erreur/NomC    [NH 6187 ]#   12 : not/Adve            0.006198     #
#  581 : erreur/NomC    [NH 6187 ]#   13 : err/Verb            0.004414     #
#  581 : erreur/NomC    [NH 6187 ]#   14 : mistaken/AdjQ       0.003276     #
#  581 : erreur/NomC    [NH 6187 ]#   15 : right/Adve          0.003122     #
#  581 : erreur/NomC    [NH 6187 ]#   16 : incorrect/AdjQ      0.003091     #
#  581 : erreur/NomC    [NH 6187 ]#   17 : flaw/NomC           0.002753     #
#  581 : erreur/NomC    [NH 6187 ]#   18 : way/NomC            0.002476     #
#  581 : erreur/NomC    [NH 6187 ]#   19 : have/Verb           0.002414     #
#  581 : erreur/NomC    [NH 6187 ]#   20 : correctly/Adve      0.002414     #
#  581 : erreur/NomC    [NH 6187 ]#   21 : bad/AdjQ           0.002199     #
#  581 : erreur/NomC    [NH 6187 ]#   22 : will/Verb          0.001967     #
#  581 : erreur/NomC    [NH 6187 ]#   23 : say/Verb           0.001890     #
#  581 : erreur/NomC    [NH 6187 ]#   24 : grievous/AdjQ      0.001763     #
#  581 : erreur/NomC    [NH 6187 ]#   25 : oversight/NomC     0.001740     #
#  581 : erreur/NomC    [NH 6187 ]#   26 : recall/Verb        0.001717     #
#  581 : erreur/NomC    [NH 6187 ]#   27 : miscarriage/NomC   0.001490     #
```

```
#  581 : erreur/NomC     [NH 6187 ]#   28 : fault/NomC          0.001451      #
#  581 : erreur/NomC     [NH 6187 ]#   29 : erroneous/AdjQ      0.001444      #
#  581 : erreur/NomC     [NH 6187 ]#   30 : in fact/Adve        0.001413      #
#  581 : erreur/NomC     [NH 6187 ]#   31 : case/NomC           0.001398      #
#  581 : erreur/NomC     [NH 6187 ]#   32 : clerical/AdjQ       0.001259      #
#  581 : erreur/NomC     [NH 6187 ]#   33 : failure/NomC        0.001251      #
#  581 : erreur/NomC     [NH 6187 ]#   34 : now/Adve            0.001151      #
#  581 : erreur/NomC     [NH 6187 ]#   35 : correct/AdjQ        0.001140      #
#  581 : erreur/NomC     [NH 6187 ]#   36 : statement/NomC      0.001048      #
#  581 : erreur/NomC     [NH 6187 ]#   37 : lead/Verb           0.001036      #
#  581 : erreur/NomC     [NH 6187 ]#   38 : memory/NomC         0.001013      #
#  581 : erreur/NomC     [NH 6187 ]#   39 : fallacy/NomC        0.000963      #
#  [...]                                                                      #
#================================#=======================================#
```

Tableau 39 : *Ligne de matrice pour le nom 'erreur'*

Les paramètres pour cette requête sont les mêmes que pour *mistake/NomC*, sauf le nom du fichier dans lequel le programme repère l'information. Le mot *erreur/NomC* apparaît 6 187 fois dans le corpus, ce qui en fait le 581^{e} mot dans la partie française du bitexte en termes de fréquence.

Le tableau pour *erreur/NomC* est tout aussi révélateur que celui de *mistake/NomC*. Les deux premiers équivalents possibles sont *mistake/NomC* et *error/NomC*, exactement ce qu'on aurait trouvé dans les dictionnaires bilingues *traditionnels*. Figurent aussi sur la liste des équivalents directs qui, dans certains contextes, pourraient effectivement être des traductions de *erreur/NomC*, notamment *flaw/NomC*, *oversight/NomC*, etc. Cette liste contient aussi des équivalents comme *mistake/Verb* et *err/Verb* qui sont des équivalents de collocations comme *commettre une erreur* et *être dans l'erreur*. Mais, comme c'était le cas pour *mistake/NomC*, ce sont les possibilités collocationnelles entre les mots qui constituent cette liste qui sont particulièrement intéressantes pour nous, par exemple, *make* + *mistake/error*, *correct* + *mistake/error* et *grievous* + *mistake*.

4.1.2 Coloc

Si *matrix_consult* permet de retrouver dans un dictionnaire bilingue probabiliste les *n* premiers équivalents possibles en L2 (langue 2) d'un mot en L1 (langue 1), *coloc*[2] utilise ces résultats pour produire les *n* premiers équivalents possibles en L1 d'un mot donné en L1.

Le principe que nous avons élaboré[3] est expliqué ici pour un mot anglais donné en entrée, mais il est réversible : on peut aussi trouver des mots en français à partir d'un mot français.

Lorsqu'on demande à *coloc* de traiter un mot anglais, il récupère d'abord la ligne de matrice anglaise sur laquelle apparaît ce mot anglais *e* et obtient une série d'équivalents possibles *f*. *Coloc* prend ensuite les 200 premiers équivalents *f* et les cherche à leur tour dans la matrice française, produisant ainsi, pour chaque mot français *f*, une série d'équivalents anglais *e'* (on ajoute le *prime* après le *e* pour faire *e'*, c'est-à-dire *e prime*, afin d'éviter toute confusion avec le mot de départ). *Coloc* calcule ensuite la probabilité de chacun de ces équivalents anglais *e'* étant donné le mot anglais initial *e*, soit $Pr(e'|e)$, en multipliant deux probabilités, soit $Pr\_1(f|e)$ de la première étape (où *coloc* va chercher la ligne de matrice pour le mot anglais initial *e*) et $Pr\_2(e'|f)$ la probabilité à la seconde étape (où *coloc* repère les équivalents anglais possibles pour le mot français trouvé à la première étape). Une fois les probabilités calculées, *coloc* fait ensuite la somme de toutes les probabilités pour chaque mot anglais *e'* (il est fort probable qu'un mot anglais soit l'équivalent

---

[2] Le nom de ce programme ne prend pas la majuscule.

[3] C'est Pierre Plamondon, du CITI, qui en a fait la programmation.

de plusieurs mots français différents) et classe les mots en ordre décroissant de probabilité ou, plus formellement,

$$c(e') = Somme\_f\ Pr\_1(f|e)\ Pr\_2(e'|f)$$

où c est la valeur dont *coloc* se sert pour trier la liste d'équivalents possibles.

Voici la sortie produite par *coloc* pour le mot *mistake/NomC* (tableau 40).

```
balzac% coloc -n 50 -nc 2
coloc> mistake/NomC
```

Mot 'mistake/NomC' Med NbElems 50 de 680 Cmb NbElems 10363

mistake/NomC	0.179707	act/NomC	0.002989
error/NomC	0.120998	repeat/Verb	0.002821
make/Verb	0.092983	offence/NomC	0.002642
commit/Verb	0.081870	then/Adve	0.002592
wrong/Adve	0.051388	incorrect/AdjQ	0.002373
be/Verb	0.042497	fool/Verb	0.002355
believe/Verb	0.036419	mistaken/AdjQ	0.002298
think/Verb	0.032089	right/Adve	0.002289
understand/Verb	0.031097	place/NomC	0.002251
mistake/Verb	0.009543	take/Verb	0.002246
serious/AdjQ	0.009189	again/Adve	0.002103
not/Adve	0.009168	bad/AdjQ	0.002090
understanding/NomC	0.007869	people/NomC	0.002048
have/Verb	0.007523	guilty/AdjQ	0.001970
correct/Verb	0.007279	involve/Verb	0.001969
wrong/AdjQ	0.007089	flaw/NomC	0.001911
do/Verb	0.006017	way/NomC	0.001886
mislead/Verb	0.005117	perpetrate/Verb	0.001863
fault/NomC	0.004621	deceive/Verb	0.001849
person/NomC	0.004332	case/NomC	0.001846
will/Verb	0.003733	deal/Verb	0.001789
occur/Verb	0.003655	let/Verb	0.001767
err/Verb	0.003414	past/NomC	0.001509
say/Verb	0.003121	correctly/Adve	0.001453
crime/NomC	0.003010	false/AdjQ	0.001378

Tableau 40 : *Sortie de coloc pour le nom 'mistake'*

Sauf indication contraire, *coloc* travaille sur l'anglais par défaut. Ainsi, il suffit de taper *coloc* sans préciser la matrice à consulter. Il faut toutefois indiquer le nombre d'équivalents à afficher (*-n 50*) et le nombre de colonnes sur lesquelles ces équivalents doivent être présentés (*-nc 2*). Il suffit ensuite, à l'invite *coloc>*, de taper le mot anglais voulu ainsi que sa catégorie grammaticale, séparés par une barre oblique, soit *mistake/NomC.*

Coloc affiche ensuite, en ordre décroissant de *c*, les *e'* qui ont été calculés à partir de *e*, précédés de quelques statistiques. Ici, par exemple, *coloc* précise qu'il n'affiche que les 50 premiers *e'* alors que la liste complète d'équivalents en comprend 680. Il indique ensuite que, pour arriver à ces résultats, il a combiné les 10 363 valeurs de *f* produites à la première étape.

Le tableau 41 présente les résultats de *coloc* pour le mot *erreur/NomC.* La description des divers éléments est la même que pour *mistake/NomC*, à une différence près : il faut préciser quelles matrices utiliser puisque les valeurs par défaut sont pour un mot où L1 est l'anglais.

```
balzac% coloc -tm all.flx.tag.rev.ml -rtm all.flx.tag.ml -n 50 -nc 1
coloc> erreur/NomC
```

Mot 'erreur/NomC' Med NbElems 50 de 1157 Cmb NbElems 7877

erreur/NomC	0.322879	mal/Adve	0.003673
commettre/Verb	0.089418	faute/NomC	0.003512
croire/Verb	0.054304	erroné/AdjQ	0.003431
comprendre/Verb	0.040734	dire/Verb	0.003425
faire/Verb	0.037253	compréhension/NomC	0.002629
avoir/Verb	0.030789	répéter/Verb	0.002588
penser/Verb	0.028928	lacune/NomC	0.002582
tromper/Verb	0.028147	abuser/Verb	0.002317
être/Verb	0.024847	glisser/Verb	0.002289
tort/NomC	0.013363	raison/NomC	0.002196
ne/Adve	0.011229	aller/Verb	0.002186
pas/Adve	0.009630	méprendre/Verb	0.002149
faux/AdjQ	0.008539	pouvoir/Verb	0.002061
mauvais/AdjQ	0.007573	contenir/Verb	0.001912
rendre/Verb	0.007289	devoir/Verb	0.001856
grave/AdjQ	0.006790	comporter/Verb	0.001740
estimer/Verb	0.006756	sembler/Verb	0.001708
corriger/Verb	0.006583	répréhensible/AdjQ	0.001696
à mon avis/Adve	0.005331	reconnaître/Verb	0.001539
bien/Adve	0.005208	bon/AdjQ	0.001451
prendre/Verb	0.004295	vouloir/Verb	0.001403
savoir/Verb	0.004157	fait/NomC	0.001378
avis/NomC	0.004012	inexact/AdjQ	0.001358
mal/NomC	0.003899	tenir/Verb	0.001339
chose/NomC	0.003884	présenter/Verb	0.001292

Tableau 41 : ***Sortie de coloc pour le nom 'erreur'***

Encore une fois, des paires de mots tirés de cette liste semblent pointer vers des collocations, par exemple *commettre* + *erreur*, *avoir* + *tort*, *grave* + *erreur*, etc.

4.1.3 Information mutuelle

Le calcul de l'information mutuelle (Church et Hanks 1990) compare la probabilité d'observer les mots *X* et *Y* ensemble (probabilité conjointe) à la probabilité de les observer ensemble simplement par l'effet du hasard. Si effectivement les mots *X* et *Y* sont associés, la probabilité de les voir ensemble, $P(x,y)$, sera plus grande que la probabilité de les observer l'un près de l'autre de façon aléatoire, $P(x)\ P(y)$. Plus les mots sont liés, plus la valeur de l'information mutuelle sera grande.

Pour y arriver, on calcule $P(x)$ et $P(y)$ en divisant leur fréquence respective par le nombre de mots *N* contenus dans le corpus (processus de normalisation). La probabilité conjointe $P(x,y)$, elle, est évaluée en comptant le nombre de fois où le mot *X* est suivi du mot *Y* dans une fenêtre de *m* mots, $f_m(x,y)$, et en normalisant encore une fois par la taille du corpus *N*. Ainsi, l'information mutuelle *I* pour le couple (x,y) se calcule comme suit :

Calcul de l'information mutuelle *I*
$$I(x,y) \equiv \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Tableau 42 : *Information mutuelle – équation*

La grandeur de la fenêtre peut être ajustée selon le type de relation à l'étude, qu'elle soit figée, composée, sémantique ou lexicale. Le tableau 43 présente quelques-uns des résultats rapportés et expliqués par Church et Hanks (1990:23).

Distance entre *x* et *y* -- Moyenne et variance				
Relation	**Mot *x***	**Mot *y***	**Distance**	
			Moyenne	**Variance**
Figée	bread drink	butter drive	2,00 2,00	0,00 0,00
Composée	computer United	scientist States	1,12 0,98	0,10 0,14
Sémantique	man man	woman women	1,46 -0,12	8,07 13,08
Lexicale	refraining coming keeping	from from from	1,11 0,83 2,14	0,20 2,89 5,53

Tableau 43 : *Information mutuelle : quelques exemples*

On remarque que *bread* et *butter* ainsi que *drink* et *drive* sont, en moyenne, à deux mots l'un de l'autre (puisque leur forme complète respective est *bread and butter* et *drink and drive*), et que cette distance ne varie jamais (variance de 0). Il est alors facile d'affirmer que nous avons affaire à des relations figées : on aura toujours *bread..mot..butter* et *drink..mot..drive*. Les composés ont aussi une structure très figée (faible variance), et les mots qui les composent sont à une

distance moyenne de 1, c'est-à-dire qu'ils sont côte à côte. C'est le cas de *computer scientist* et de *United States*.

En revanche, la relation sémantique qui lie *man* et *woman* est beaucoup plus souple comme l'indique la variance (8,07). Le cas de *(man,women)*, où la distance moyenne est très près de 0, indique que *women* se place avant ou après *man* dans une proportion à peu près égale, mais sans doute un peu plus souvent devant puisque la valeur est négative.

Les relations lexicales, qui sont particulièrement nombreuses, se prêtent mal à la généralisation. Par exemple, la combinaison *refraining...from* semble assez figée : ces deux mots sont à une distance de 1,11 mot, et la variance est faible. Toutefois, le couple *keeping...from* est séparé, en moyenne, par un mot, ou à une distance de 2 (probablement un complément d'objet direct comme *keep someone from...*), et sa variance est assez élevée, ce qui signifie que la structure de ce couple est très peu figée.

L'exemple de *save* (tableau 44), tiré de Church et Hanks (1990:27), illustre bien en quoi l'information mutuelle est intéressante pour le repérage des collocations et pour la lexicographie en général.

Mots apparaissant souvent à la droite de *save*					
I(x,y)[4]	f(x,y)	f(x)	x	f(y)	y
9,5	6	724	save	170	forests
9,4	6	724	save	180	$1.2
8,8	37	724	save	1697	lives
8,7	6	724	save	301	enormous
8,3	7	724	save	447	annually
7,7	20	724	save	2001	jobs
7,6	64	724	save	6776	money
7,2	36	724	save	4875	life
6,6	8	724	save	1668	dollars
6,4	7	724	save	1719	costs
6,4	6	724	save	1481	thousands
6,2	9	724	save	2590	face
5,7	6	724	save	2311	son
5,7	6	724	save	2387	estimated
5,5	7	724	save	3171	your

Tableau 44 : *Information mutuelle – exemple de 'save'*

Cette liste montre qu'on peut utiliser *save* au sens d'économiser avec *money*, *dollars* et *cost*, au sens de protéger avec *forests*, *jobs*. Les calculs d'information mutuelle ont aussi repéré l'expression figée *to save face*.

[4] I = calcul d'information mutuelle
f(x,y) = nombre d'occurrences de x et de y ensemble, dans cet ordre
f(x) = nombre d'occurrences de x
f(y) = nombre d'occurrences de y

4.2 Méthodologie de l'extraction automatique des collocations

Il est évident, après avoir examiné les sorties de *matrix_consult* et de *coloc*, que l'un et l'autre modèle livrent des résultats qui, traités avec les calculs de l'information mutuelle, ouvrent la voie à l'extraction automatique de collocations dans des bitextes. Pour un mot donné en L1, chacun des deux modèles donne une liste de collocations potentielles, une en L1 et l'autre en L2. Il faut maintenant décider d'une stratégie qui produira facilement des résultats utilisables dans le cadre de la lexicographie bilingue.

4.2.1 Étude de cas : erreur/NomC

La présente section décrit, étape par étape, la méthodologie suivie pour extraire des listes de collocations pour le mot *erreur/NomC*.

4.2.1.1 Extraction des collocations de erreur/NomC en L1

Coloc a été expliqué en détail à la section précédente. Rappelons cependant que, pour un mot donné en L1, *coloc* passe par L2 et revient vers L1 pour dresser une liste des collocatifs possibles en L1. De façon générale, la liste L1-1 est affichée comme suit, à raison d'un collocatif par ligne :

Forme générale de la liste L1-1
mot_1/cat_gram *mot_2/cat_gram* *mot_3/cat_gram* (...) *mot_n/cat_gram*

Tableau 45 : Coloc – forme générale de la liste L1-1

Les pages suivantes (tableau 46) montrent la sortie de *coloc* pour le mot à l'étude, soit *erreur/NomC*.

Liste L1-1 pour *erreur/NomC*

erreur/NomC	0.322879	permettre/Verb	0.001003
commettre/Verb	0.089418	tout à fait/Adve	0.000952
croire/Verb	0.054304	falloir|faillir/Verb	0.000919
comprendre/Verb	0.040734	là/Adve	0.000901
faire/Verb	0.037253	suivre|être/Verb	0.000892
avoir/Verb	0.030789	exact/AdjQ	0.000865
penser/Verb	0.028928	remédier/Verb	0.000848
tromper/Verb	0.028147	déclaration/NomC	0.000848
être/Verb	0.024847	leurrer/Verb	0.000846
tort/NomC	0.013363	mémoire/NomC	0.000838
ne/Adve	0.011229	entente/NomC	0.000821
pas/Adve	0.009630	clocher/Verb	0.000814
faux/AdjQ	0.008539	convaincre/Verb	0.000792
mauvais/AdjQ	0.007573	donner/Verb	0.000781
rendre/Verb	0.007289	réparer/Verb	0.000762
grave/AdjQ	0.006790	à tort/Adve	0.000751
estimer/Verb	0.006756	maintenant/Adve	0.000743
corriger/Verb	0.006583	entendre/Verb	0.000739
à mon avis/Adve	0.005331	considérer/Verb	0.000729
bien/Adve	0.005208	relever/Verb	0.000688
prendre/Verb	0.004295	mieux/Adve	0.000684
savoir/Verb	0.004157	apporter/Verb	0.000668
avis/NomC	0.004012	député/NomC	0.000653
mal/NomC	0.003899	très/Adve	0.000652
chose/NomC	0.003884	doute/NomC	0.000649
mal/Adve	0.003673	malheureusement/Adve	0.000648
faute/NomC	0.003512	réfléchir/Verb	0.000645
erroné/AdjQ	0.003431	ministre/NomC	0.000639
dire/Verb	0.003425	sujet/NomC	0.000636
compréhension/NomC	0.002629	produire/Verb	0.000634
répéter/Verb	0.002588	attribuable/AdjQ	0.000627
lacune/NomC	0.002582	risque/NomC	0.000613
abuser/Verb	0.002317	juger/Verb	0.000602
glisser/Verb	0.002289	tourner/Verb	0.000595
raison/NomC	0.002196	songer/Verb	0.000593
aller/Verb	0.002186	correct/AdjQ	0.000591
méprendre/Verb	0.002149	injuste/AdjQ	0.000584
pouvoir/Verb	0.002061	vrai/AdjQ	0.000581
contenir/Verb	0.001912	imaginer/Verb	0.000575
devoir/Verb	0.001856	gouvernement/NomC	0.000573
comporter/Verb	0.001740	défaut/NomC	0.000562
sembler/Verb	0.001708	laisser/Verb	0.000560
répréhensible/AdjQ	0.001696	saisir/Verb	0.000548
reconnaître/Verb	0.001539	complètement/Adve	0.000543
bon/AdjQ	0.001451	pire/AdjQ	0.000536
vouloir/Verb	0.001403	sens/NomC	0.000527
fait/NomC	0.001378	monde/NomC	0.000524
inexact/AdjQ	0.001358	douter/Verb	0.000517
tenir/Verb	0.001339	mesure/NomC	0.000517
présenter/Verb	0.001292	simplement/Adve	0.000508
monumental/AdjQ	0.001281	mettre/Verb	0.000502
agir/Verb	0.001253	plus/Adve	0.000501
trouver/Verb	0.001246	assurer/Verb	0.000498
route/NomC	0.001233	situation/NomC	0.000495
rappeler/Verb	0.001173	convenir/Verb	0.000483
façon/NomC	0.001152	là-dessus/Adve	0.000480
bévue/NomC	0.001138	judiciaire/AdjQ	0.000477
rectifier/Verb	0.001093	clair/AdjQ	0.000475
en fait/Adve	0.001049	sans doute/Adve	0.000472
souvenir/Verb	0.001044	passé/NomC	0.000470
éviter/Verb	0.001043	reprocher/Verb	0.000468
cas/NomC	0.001034	venir/Verb	0.000467
arriver/Verb	0.001006	gagner/Verb	0.000460

Liste L1-1 pour *erreur/NomC* (suite)

possibilité/NomC	0.000457	grand/AdjQ	0.000399
effectuer/Verb	0.000456	appeler/Verb	0.000396
tout/Adve	0.000452	idée/NomC	0.000395
chambre/NomC	0.000450	capable/AdjQ	0.000394
voir/Verb	0.000449	point/NomC	0.000393
passé/AdjQ	0.000445	encore/Adve	0.000392
prétendre/Verb	0.000442	compte/NomC	0.000386
prouver/Verb	0.000441	partie/NomC	0.000386
inacceptable/AdjQ	0.000440	échec/NomC	0.000384
alors/Adve	0.000439	première|premier/NomC	0.000383
rond/AdjQ	0.000437	moyen/NomC	0.000376
aussi/Adve	0.000429	dernier/AdjQ	0.000363
connaissance/NomC	0.000427	juste/AdjQ	0.000358
oubli/NomC	0.000425	banque/NomC	0.000329
valoir/Verb	0.000417	servir/Verb	0.000327
même/Adve	0.000415	incorrect/AdjQ	0.000326
conséquence/NomC	0.000412	à l'occasion/NomC	0.000290
terrible/AdjQ	0.000412	déclarer/Verb	0.000290
redire/Verb	0.000409	réaliser/Verb	0.000290
paraître/Verb	0.000408	contexte/NomC	0.000277
peut-être/Adve	0.000408	accumuler/Verb	0.000276
signaler/Verb	0.000402	correction/NomC	0.000275

Tableau 46 : *Coloc – liste L1-1 pour le nom 'erreur'*

L'analyse de cette liste révèle certains mots qui pourraient être collocatifs d'*erreur*. À titre d'exemples, citons *commettre (une erreur)*, *grave (erreur)*, *corriger (une erreur)*, *glisser (une erreur)* et *(erreur) monumental(e)*.

Avant de soumettre L1-1 à l'étape suivante, soit le tri par information mutuelle pour repérer, justement, ces combinaisons, elle doit être 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, c'est-à-dire lorsque le dictionnaire morphologique ne pouvait

distinguer entre deux formes semblables de mots différents (*suis*, par exemple, qui pourrait être soit le verbe *être* ou le verbe *suivre*), certains *mots* apparaissent dans L1-1 sous forme de doublets. Ainsi, afin d'optimiser les résultats du tri par information mutuelle, les mots très vagues sont éliminés de L1-1 et les doublets, séparés.

Le tableau 47 présente, pour *erreur/NomC*, les mots que nous avons exclus (*stop-word list* en anglais) de l'étape suivante (calcul de l'information mutuelle) et des doublets qu'il s'est avéré nécessaire de séparer. Pour l'instant, cette opération se fait manuellement.

Modifications apportées à L1-1 pour *erreur/NomC*		
Mots exclus (ordre alphabétique)		**Doublets divisés**
à l'occasion	là-dedans	falloir|faillir/Verb
à mon avis	malheureusement	première|premier/NomC
à tout	même	suivre|être/Verb
alors	ne	
aussi	pas	
avoir	peut-être	
bien	plus	
en fait	sans doute	
encore	tout	
être	tout à fait	
là	très	

Tableau 47 : *Coloc – liste des mots exclus (L1-1) pour 'erreur'*

De plus, nous avons supprimé, avec la commande *cut* en Unix, la probabilité associée aux membres de la liste L1-1 pour former la liste simplifiée L1-2, qui ne contient que des

mot/cat_gram puisque le calcul de l'information mutuelle ne prend en entrée que des chaînes du type *mot/cat_gram*. C'est cette liste qui sera traitée à l'étape suivante, étape qui consiste à prendre L1-2 et à la multiplier par elle-même (produit cartésien) pour produire des paires de mots (tableau 48) :

Produit cartésien de L1-2 x L1-2[5]					
	mot_1	**mot_2**	**mot_3**	...	**mot_i**
mot_1	(mot_1, mot_1)	(mot_1, mot_2)	(mot_1, mot_3)		(mot_1, mot_i)
mot_2	(mot_2, mot_1)	(mot_2, mot_2)	(mot_2, mot_3)		(mot_2, mot_i)
mot_3	(mot_3, mot_1)	(mot_3, mot_2)	(mot_3, mot_3)		(mot_3, mot_i)
...					
mot_i	(mot_i, mot_1)	(mot_i, mot_2)	(mot_i, mot_3)		(mot_i, mot_i)

Tableau 48 : *Coloc — produit cartésien L1-2 x L1-2*

À chacune de ces paires, *mtf_consult* (l'algorithme qui effectue le calcul de l'information mutuelle) associe une valeur[6]. Une fois les paires triées en fonction de cette valeur, elles constituent la liste L1-3. Le tableau 49 présente les 100 premières combinaisons calculées de cette façon à partir du mot *erreur/NomC*.

---

[5] 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).

[6] Puisque nous tentons toujours de déterminer la meilleure façon d'effectuer le calcul, nous ne décrirons pas, dans le cadre du présent travail, les calculs effectués pour obtenir les résultats actuels.

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

député/NomC	pouvoir/Verb	27179.203125
pouvoir/Verb	député/NomC	14047.737305
croire/Verb	député/NomC	7828.003418
monsieur/NomC	façon/NomC	5576.541016
député/NomC	croire/Verb	5561.788086
pouvoir/Verb	compte/NomC	2417.564209
tourner/Verb	rond/AdjQ	2250.385010
compte/NomC	pouvoir/Verb	2052.706543
cas/NomC	question/NomC	1616.506226
question/NomC	cas/NomC	1370.025024
gouvernement/NomC	mesure/NomC	1311.900635
gouvernement/NomC	mettre/Verb	1286.209473
bévue/NomC	monumental/AdjQ	881.220093
mesure/NomC	gouvernement/NomC	865.073181
mémoire/NomC	fidèle/AdjQ	590.637573
façon/NomC	monsieur/NomC	564.110779
mettre/Verb	gouvernement/NomC	471.030365
bévue/NomC	rectifier/Verb	469.761932
aller/Verb	tenir/Verb	404.533875
échec/NomC	monumental/AdjQ	399.188843
commettre/Verb	bévue/NomC	389.679321
tenir/Verb	aller/Verb	368.697815
faire/Verb	comprendre/Verb	329.775604
devoir/Verb	faire/Verb	315.242920
savoir/Verb	faire/Verb	314.913605
comprendre/Verb	faire/Verb	289.721497
prendre/Verb	mesure/NomC	275.878052
inexact/AdjQ	incorrect/AdjQ	275.605255
erreur/NomC	monumental/AdjQ	271.444000
question/NomC	gouvernement/NomC	260.903625
ministre/NomC	penser/Verb	256.617859
gouvernement/NomC	question/NomC	255.101212
bévue/NomC	passé/AdjQ	254.614822
rectifier/Verb	oubli/NomC	238.731461
monsieur/NomC	faire/Verb	231.840179
comporter/Verb	lacune/NomC	228.613922
penser/Verb	ministre/NomC	226.021133
comporter/Verb	faille/NomC	223.840881
ministre/NomC	laisser/Verb	219.224854
dire/Verb	droit/NomC	215.064346
question/NomC	ministre/NomC	211.225571
juger/Verb	correct/AdjQ	207.114883
réparer/Verb	tort/NomC	197.935425
correct/AdjQ	fautif/AdjQ	197.344177
bévue/NomC	commettre/Verb	194.839661
mesure/NomC	prendre/Verb	193.614304
dernier/AdjQ	gouvernement/NomC	192.880142
député/NomC	question/NomC	188.444183
réparer/Verb	oubli/NomC	186.157074
ministre/NomC	gouvernement/NomC	183.117691

Liste L1-3 pour *erreur/NomC* (suite)		
répréhensible/AdjQ	incorrect/AdjQ	180.075409
faire/Verb	savoir/Verb	179.745651
commettre/Verb	erreur/NomC	178.145248
dire/Verb	donner/Verb	169.940094
penser/Verb	chambre/NomC	168.694611
gouvernement/NomC	dernier/AdjQ	164.783264
apporter/Verb	correction/NomC	163.715698
ministre/NomC	question/NomC	162.006241
gouvernement/NomC	pouvoir/Verb	161.950790
monsieur/NomC	bon/AdjQ	161.840942
corriger/Verb	lacune/NomC	160.700623
erreur/NomC	glisser/Verb	160.140594
ministre/NomC	concerner/Verb	156.821350
commettre/Verb	monumental/AdjQ	146.870010
devoir/Verb	bon/AdjQ	146.610443
réparer/Verb	bévue/NomC	146.523621
gouvernement/NomC	suivre/Verb	143.390854
fautif/AdjQ	inexact/AdjQ	140.986191
correction/NomC	incorrect/AdjQ	140.795502
député/NomC	sujet/NomC	139.177368
simplement/Adve	gouvernement/NomC	138.299011
répréhensible/AdjQ	fautif/AdjQ	138.176682
député/NomC	dire/Verb	135.705002
dire/Verb	peut-être/Adve	135.190704
député/NomC	grand/AdjQ	133.108292
éviter/Verb	malentendu/NomC	133.071747
glisser/Verb	oubli/NomC	132.101425
oubli/NomC	glisser/Verb	132.101425
trouver/Verb	redire/Verb	131.874985
ministre/NomC	faire/Verb	130.280167
rectifier/Verb	erroné/AdjQ	125.901619
pouvoir/Verb	faire/Verb	125.837425
question/NomC	mesure/NomC	125.420639
devoir/Verb	rendre/Verb	123.199776
inexact/AdjQ	erroné/AdjQ	123.108978
pouvoir/Verb	gouvernement/NomC	122.908707
ministre/NomC	venir/Verb	122.733177
gouvernement/NomC	simplement/Adve	122.539986
monsieur/NomC	pouvoir/Verb	122.272552
faire/Verb	devoir/Verb	121.701233
bévue/NomC	faute/NomC	121.577332
question/NomC	député/NomC	121.436516
dire/Verb	partie/NomC	119.388519
corriger/Verb	erreur/NomC	118.084663
bon/AdjQ	devoir/Verb	117.154465
voir/Verb	gouvernement/NomC	116.241447
erreur/NomC	commettre/Verb	115.032295
devoir/Verb	pouvoir/Verb	114.081406
remédier/Verb	lacune/NomC	112.981308
reconnaître/Verb	pouvoir/Verb	110.228134

Tableau 49 : ***Coloc – liste L1-3 pour le nom 'erreur'***

Certes, cette liste est intéressante, mais, puisqu'en lexicographie, on travaille sur un mot donné, en l'occurrence *erreur/NomC*, il ne sera utile d'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 le montre le tableau 50.

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

erreur/NomC	monumental/AdjQ	271.444000
commettre/Verb	erreur/NomC	178.145248
erreur/NomC	glisser/Verb	160.140594
corriger/Verb	erreur/NomC	118.084663
erreur/NomC	commettre/Verb	115.032295
rectifier/Verb	erreur/NomC	98.477562
erreur/NomC	passé/AdjQ	76.250916
bévue/NomC	erreur/NomC	68.590683

Tableau 50 : *Coloc – liste L1-4 pour le nom 'erreur'*

Si le tableau semble 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.

4.2.1.2 Extraction de collocations en L2 associées à erreur/NomC

Si l'utilisation de *coloc* suivie du calcul d'*information mutuelle* permet d'extraire des collocations possibles d'*erreur/NomC*, ce qui suit décrit de quelle façon on peut obtenir une liste de collocations en L2 qui pourraient être des équivalents des collocations en L1 déjà extraites.

Comme c'était le cas dans la section précédente, l'explication est donnée pour *erreur/NomC*, mais elle peut être généralisée pour tout autre mot.

Dans un premier temps, il faut obtenir, avec *matrix_consult* (cela est expliqué en détail à la section 4.1.1), la liste L2-1 des traductions possibles d'*erreur/NomC* telles que trouvées dans le dictionnaire bilingue probabiliste (tableau 51).

Liste L2-1 pour *erreur/NomC*

581	erreur/NomC	[NH 6187]	0	mistake/NomC	0.286417
581	erreur/NomC	[NH 6187]	1	error/NomC	0.199803
581	erreur/NomC	[NH 6187]	2	make/Verb	0.101009
581	erreur/NomC	[NH 6187]	3	wrong/Adve	0.061885
581	erreur/NomC	[NH 6187]	4	believe/Verb	0.059670
581	erreur/NomC	[NH 6187]	5	think/Verb	0.051550
581	erreur/NomC	[NH 6187]	6	understand/Verb	0.051304
581	erreur/NomC	[NH 6187]	7	be/Verb	0.030635
581	erreur/NomC	[NH 6187]	8	mistake/Verb	0.013057
581	erreur/NomC	[NH 6187]	9	understanding/NomC	0.012872
581	erreur/NomC	[NH 6187]	10	wrong/AdjQ	0.008505
581	erreur/NomC	[NH 6187]	11	correct/Verb	0.006967
581	erreur/NomC	[NH 6187]	12	not/Adve	0.006198
581	erreur/NomC	[NH 6187]	13	err/Verb	0.004414
581	erreur/NomC	[NH 6187]	14	mistaken/AdjQ	0.003276
581	erreur/NomC	[NH 6187]	15	right/Adve	0.003122
581	erreur/NomC	[NH 6187]	16	incorrect/AdjQ	0.003091
581	erreur/NomC	[NH 6187]	17	flaw/NomC	0.002753
581	erreur/NomC	[NH 6187]	18	way/NomC	0.002476
581	erreur/NomC	[NH 6187]	19	have/Verb	0.002414
581	erreur/NomC	[NH 6187]	20	correctly/Adve	0.002414

```
#  581 : erreur/NomC    [NH 6187 ]#   21 : bad/AdjQ              0.002199    #
#  581 : erreur/NomC    [NH 6187 ]#   22 : will/Verb             0.001967    #
#  581 : erreur/NomC    [NH 6187 ]#   23 : say/Verb              0.001890    #
#  581 : erreur/NomC    [NH 6187 ]#   24 : grievous/AdjQ         0.001763    #
#  581 : erreur/NomC    [NH 6187 ]#   25 : oversight/NomC        0.001740    #
#  581 : erreur/NomC    [NH 6187 ]#   26 : recall/Verb           0.001717    #
#  581 : erreur/NomC    [NH 6187 ]#   27 : miscarriage/NomC      0.001490    #
#  581 : erreur/NomC    [NH 6187 ]#   28 : fault/NomC            0.001451    #
#  581 : erreur/NomC    [NH 6187 ]#   29 : erroneous/AdjQ        0.001444    #
#  581 : erreur/NomC    [NH 6187 ]#   30 : in fact/Adve          0.001413    #
#  581 : erreur/NomC    [NH 6187 ]#   31 : case/NomC             0.001398    #
#  581 : erreur/NomC    [NH 6187 ]#   32 : clerical/AdjQ         0.001259    #
#  581 : erreur/NomC    [NH 6187 ]#   33 : failure/NomC          0.001251    #
#  581 : erreur/NomC    [NH 6187 ]#   34 : now/Adve              0.001151    #
#  581 : erreur/NomC    [NH 6187 ]#   35 : correct/AdjQ          0.001140    #
#  581 : erreur/NomC    [NH 6187 ]#   36 : statement/NomC        0.001048    #
#  581 : erreur/NomC    [NH 6187 ]#   37 : lead/Verb             0.001036    #
#  581 : erreur/NomC    [NH 6187 ]#   38 : memory/NomC           0.001013    #
#  581 : erreur/NomC    [NH 6187 ]#   39 : fallacy/NomC          0.000963    #
#  581 : erreur/NomC    [NH 6187 ]#   40 : inaccurate/AdjQ       0.000932    #
#  581 : erreur/NomC    [NH 6187 ]#   41 : remember/Verb         0.000905    #
#  581 : erreur/NomC    [NH 6187 ]#   42 : fact/NomC             0.000905    #
#  581 : erreur/NomC    [NH 6187 ]#   43 : gather/Verb           0.000902    #
#  581 : erreur/NomC    [NH 6187 ]#   44 : egregious/AdjQ        0.000894    #
#  581 : erreur/NomC    [NH 6187 ]#   45 : repeat/Verb           0.000867    #
#  581 : erreur/NomC    [NH 6187 ]#   46 : factual/AdjQ          0.000790    #
#  581 : erreur/NomC    [NH 6187 ]#   47 : misconception/NomC    0.000771    #
#  581 : erreur/NomC    [NH 6187 ]#   48 : prove/Verb            0.000759    #
#  581 : erreur/NomC    [NH 6187 ]#   49 : opinion/NomC          0.000755    #
#  581 : erreur/NomC    [NH 6187 ]#   50 : stand/Verb            0.000752    #
#  581 : erreur/NomC    [NH 6187 ]#   51 : miscalculation/NomC   0.000740    #
#  581 : erreur/NomC    [NH 6187 ]#   52 : erroneously/Adve      0.000732    #
#  581 : erreur/NomC    [NH 6187 ]#   53 : drafting/NomC         0.000729    #
#  581 : erreur/NomC    [NH 6187 ]#   54 : misunderstanding/NomC 0.000709    #
#  581 : erreur/NomC    [NH 6187 ]#   55 : terrible/AdjQ         0.000706    #
#  581 : erreur/NomC    [NH 6187 ]#   56 : part/NomC             0.000698    #
#  581 : erreur/NomC    [NH 6187 ]#   57 : do/Verb               0.000694    #
#  581 : erreur/NomC    [NH 6187 ]#   58 : back/Adve             0.000690    #
#  581 : erreur/NomC    [NH 6187 ]#   59 : somewhere/Adve        0.000602    #
#  581 : erreur/NomC    [NH 6187 ]#   60 : wrong/NomC            0.000598    #
#  581 : erreur/NomC    [NH 6187 ]#   61 : past/NomC             0.000590    #
#  581 : erreur/NomC    [NH 6187 ]#   62 : miss/Verb             0.000571    #
#  581 : erreur/NomC    [NH 6187 ]#   63 : minister/NomC         0.000567    #
#  581 : erreur/NomC    [NH 6187 ]#   64 : accurate/AdjQ         0.000552    #
#  581 : erreur/NomC    [NH 6187 ]#   65 : track/NomC            0.000544    #
#  581 : erreur/NomC    [NH 6187 ]#   66 : occur/Verb            0.000536    #
#  581 : erreur/NomC    [NH 6187 ]#   67 : simply/Adve           0.000525    #
#  581 : erreur/NomC    [NH 6187 ]#   68 : possibility/NomC      0.000525    #
#  581 : erreur/NomC    [NH 6187 ]#   69 : false/AdjQ            0.000521    #
#  581 : erreur/NomC    [NH 6187 ]#   70 : unfortunately/Adve    0.000502    #
#  581 : erreur/NomC    [NH 6187 ]#   71 : then/Adve             0.000498    #
#  581 : erreur/NomC    [NH 6187 ]#   72 : out/Adve              0.000490    #
#  581 : erreur/NomC    [NH 6187 ]#   73 : proposition/NomC      0.000467    #
#  581 : erreur/NomC    [NH 6187 ]#   74 : serve/Verb            0.000467    #
#  581 : erreur/NomC    [NH 6187 ]#   75 : true/Adve             0.000463    #
#  581 : erreur/NomC    [NH 6187 ]#   76 : context/NomC          0.000459    #
#  581 : erreur/NomC    [NH 6187 ]#   77 : big/AdjQ              0.000456    #
#  581 : erreur/NomC    [NH 6187 ]#   78 : move/NomC             0.000444    #
#  581 : erreur/NomC    [NH 6187 ]#   79 : submission/NomC       0.000444    #
#  581 : erreur/NomC    [NH 6187 ]#   80 : may/Verb              0.000436    #
#  581 : erreur/NomC    [NH 6187 ]#   81 : point/NomC            0.000425    #
#  581 : erreur/NomC    [NH 6187 ]#   82 : flawed/AdjQ           0.000402    #
#  581 : erreur/NomC    [NH 6187 ]#   83 : argue/Verb            0.000390    #
#  581 : erreur/NomC    [NH 6187 ]#   84 : house/NomC            0.000379    #
```

```
#  581 : erreur/NomC     [NH 6187 ]#   85 : piece/NomC            0.000371    #
#  581 : erreur/NomC     [NH 6187 ]#   86 : completely/Adve       0.000359    #
#  581 : erreur/NomC     [NH 6187 ]#   87 : bank/NomC             0.000359    #
#  581 : erreur/NomC     [NH 6187 ]#   88 : somehow/Adve          0.000359    #
#  581 : erreur/NomC     [NH 6187 ]#   89 : mentality/NomC        0.000356    #
#  581 : erreur/NomC     [NH 6187 ]#   90 : innocent/AdjQ         0.000352    #
#  581 : erreur/NomC     [NH 6187 ]#   91 : you're/Verb           0.000348    #
#  581 : erreur/NomC     [NH 6187 ]#   92 : blunder/NomC          0.000344    #
#  581 : erreur/NomC     [NH 6187 ]#   93 : thing/NomC            0.000344    #
#  581 : erreur/NomC     [NH 6187 ]#   94 : guess/Verb            0.000340    #
#  581 : erreur/NomC     [NH 6187 ]#   95 : admit/Verb            0.000313    #
#  581 : erreur/NomC     [NH 6187 ]#   96 : place/NomC            0.000306    #
#  581 : erreur/NomC     [NH 6187 ]#   97 : suggest/Verb          0.000290    #
#  581 : erreur/NomC     [NH 6187 ]#   98 : indeed/Adve           0.000286    #
#  581 : erreur/NomC     [NH 6187 ]#   99 : inadvertently/Adve    0.000286    #
#  [...]                                                                      #
#=================================#=========================================#
```

Tableau 51 : *Coloc -- liste L2-1 pour le nom 'erreur'*

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, 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 manuellement les mots trop vagues ou trop

généraux de L2-1 avant de passer au produit cartésien de cette liste par elle-même[7], puis au calcul de l'information mutuelle.

Le tableau 52 présente les mots qui ont été exclus de L2-1 pour optimiser le calcul de l'information mutuelle.

Mots exclus de la liste L2-1 pour *erreur/NomC* avant le calcul de l'information mutuelle	
back	out
be	simply
do	somehow
have	somewhere
in fact	then
indeed	unfortunately
may	will
not	you're
now	

Tableau 52 : *Coloc -- liste des mots exclus (L2-1) pour 'erreur'*

Une fois de plus, la commande Unix *cut* est utilisée pour supprimer de la liste toute l'information superflue. La nouvelle liste L2-2, qui ne contient qu'une série de *mot/cat_gram*, sera l'entrée de l'étape information mutuelle.

---

[7] Ce que nous entendons par « le produit cartésien d'une liste par elle-même » est expliqué en détail à la sous-section 4.2.1.1.

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 (déjà calculée et que *mtf_consult* permet de repérer), comme le montre la liste L2-3 pour *erreur/NomC* (tableau 53).

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

goof/NomC	goof/Verb	14860.539062
egregious/AdjQ	goof/NomC	14447.746094
think/Verb	government/NomC	3604.736084
egregious/AdjQ	error/NomC	2970.652832
fallacy/NomC	mistakenly/Adve	2110.872070
government/NomC	pass/Verb	1807.241455
sin/NomC	omission/NomC	1601.758179
government/NomC	think/Verb	1294.454346
mistake/Verb	untrue/NomC	1215.988647
grievous/AdjQ	error/NomC	1208.401123
pass/Verb	government/NomC	1103.818237
inadvertently/Adve	incorrectly/Adve	923.287964
mistakenly/Adve	execute/Verb	915.057800
quasi-judicial/AdjQ	body/NomC	895.561707
clerical/AdjQ	error/NomC	894.111389
factual/AdjQ	erroneously/Adve	821.240295
sin/NomC	time-consuming/AdjQ	758.928284
inadvertently/Adve	mislead/Verb	701.133850
think/Verb	people/NomC	689.264465
gather/Verb	glitch/NomC	651.288330
erroneously/Adve	inadvertently/Adve	631.723328
inadvertently/Adve	erroneously/Adve	631.723328
fallacy/NomC	misconception/NomC	603.106323
presume/Verb	goof/Verb	596.010925
drafting/NomC	oversight/NomC	561.371704
mark/NomC	goof/NomC	556.276855
innocent/AdjQ	execute/Verb	509.713867
erroneously/Adve	execute/Verb	505.689850
detect/Verb	wrong/NomC	482.798553
detect/Verb	wrong/NomC	482.798553
wrong/NomC	quasi-judicial/AdjQ	476.910767
error/NomC	omission/NomC	460.632477
memory/NomC	correctly/Adve	439.915405
execute/Verb	innocent/AdjQ	407.771118
incorrectly/Adve	conceive/Verb	387.685486
goof/Verb	okay/Adve	386.993195
okay/Adve	goof/Verb	386.993195
minister/NomC	make/Verb	383.594574
big/AdjQ	goof/NomC	346.284180
conceive/Verb	blunder/NomC	343.630310
goof/NomC	error/NomC	339.503174
misconception/NomC	mistaken/AdjQ	331.708466
innocent/AdjQ	mistakenly/Adve	328.482300
erroneous/AdjQ	error/NomC	305.989960
do/Verb	statement/NomC	293.560028
people/NomC	think/Verb	283.564148
factual/AdjQ	error/NomC	264.812500
guess/Verb	erratic/AdjQ	261.760864
error/NomC	grievous/AdjQ	258.943085
correct/Verb	wrong/NomC	248.559418

Liste L2-3 pour *erreur/NomC* (suite)		
mistake/NomC	blunder/NomC	247.941010
think/Verb	make/Verb	246.429367
false/AdjQ	untrue/NomC	243.683884
oversight/NomC	drafting/NomC	240.587860
paper/NomC	clerical/AdjQ	239.775208
drafting/NomC	error/NomC	238.493958
statement/NomC	do/Verb	234.188324
again/Adve	member/NomC	229.942947
sin/NomC	wrong/NomC	228.249115
untrue/NomC	mislead/Verb	226.059998
err/Verb	side/NomC	223.202728
clerical/AdjQ	oversight/NomC	222.222366
right/Adve	wrong/NomC	221.493393
presume/Verb	innocent/AdjQ	221.330322
dead/Adve	wrong/Adve	216.449921
incorrectly/Adve	exact/AdjQ	215.566513
member/NomC	again/Adve	210.468948
inadvertent/AdjQ	error/NomC	209.282776
goof/NomC	mistake/NomC	207.798187
mistake/NomC	miscalculation/NomC	203.280838
inadvertent/AdjQ	mislead/Verb	198.189606
say/Verb	go/Verb	190.232071
correct/Verb	error/NomC	189.891602
mislead/Verb	inaccurate/AdjQ	188.476837
mistake/NomC	goof/Verb	178.112732
flawed/AdjQ	erroneous/AdjQ	174.093842
egregious/AdjQ	mistake/NomC	173.165146
correction/NomC	offender/NomC	171.204361
inadvertent/AdjQ	mistake/NomC	170.793030
true/Adve	untrue/NomC	170.582977
inadvertently/Adve	misunderstand/Verb	163.859970
totally/Adve	inaccurate/AdjQ	160.983200
honest/AdjQ	mistaken/AdjQ	160.222824
accurate/AdjQ	factual/AdjQ	159.627274
grievous/AdjQ	mistake/NomC	158.490143
totally/Adve	erroneous/AdjQ	157.068024
mistakenly/Adve	accident/NomC	153.644958
mistaken/AdjQ	mentality/NomC	152.856247
error/NomC	oversight/NomC	152.016357
erroneous/AdjQ	incorrect/AdjQ	147.832443
make/Verb	minister/NomC	146.773102
detect/Verb	shortcoming/NomC	146.603210
detect/Verb	shortcoming/NomC	146.603210
wrong/NomC	execute/Verb	144.482803
grievous/AdjQ	terrible/AdjQ	143.498215
terrible/AdjQ	grievous/AdjQ	143.498215
erratic/AdjQ	occur/Verb	143.046997
presume/Verb	inadvertent/AdjQ	142.879333
error/NomC	egregious/AdjQ	141.459656
dead/Adve	track/NomC	140.245850

Tableau 53 : *Coloc – liste L2-3 pour le nom 'erreur'*

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 manuellement autour des substantifs, puisque le mot *erreur*, qui fait l'objet de cette étude, est un substantif (tableau 54).

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

```
error
     egregious/AdjQ      error/NomC 2970.652832 |
      grievous/AdjQ      error/NomC 1208.401123 |
       clerical/AdjQ      error/NomC 894.111389 |
          error/NomC   omission/NomC 460.632477 |
           goof/NomC      error/NomC 339.503174 |
     erroneous/AdjQ      error/NomC 305.989960 |
       factual/AdjQ      error/NomC 264.812500 |
         error/NomC   grievous/AdjQ 258.943085 |
      drafting/NomC      error/NomC 238.493958 |
   inadvertent/AdjQ      error/NomC 209.282776 |
       correct/Verb      error/NomC 189.891602 |
         error/NomC  oversight/NomC 152.016357 |
         error/NomC  egregious/AdjQ 141.459656 |
 misconception/NomC      error/NomC 138.572723 |
        error/NomC erroneously/Adve 134.014404 |
        error/NomC incorrectly/Adve 130.578140 |
         error/NomC    blunder/NomC 115.739723 |
         error/NomC   drafting/NomC 112.232452 |
     error/NomC miscalculation/NomC 110.707558 |
         error/NomC  erroneous/AdjQ 109.282143 |
          error/NomC    execute/Verb 87.802544 |
mistake
       mistake/Verb     untrue/NomC 1215.988647 |
        mistake/NomC    blunder/NomC 247.941010 |
           goof/NomC    mistake/NomC 207.798187 |
   mistake/NomC miscalculation/NomC 203.280838 |
       mistake/NomC       goof/Verb 178.112732 |
     egregious/AdjQ    mistake/NomC 173.165146 |
   inadvertent/AdjQ    mistake/NomC 170.793030 |
      grievous/AdjQ    mistake/NomC 158.490143 |
      mistake/NomC inadvertent/AdjQ 128.094772 |
       mistake/Verb  misguided/AdjQ 102.940834 |
flaw
        correct/Verb       flaw/NomC 137.441940 |
oversight
       drafting/NomC  oversight/NomC 561.371704 |
      oversight/NomC   drafting/NomC 240.587860 |
       clerical/AdjQ  oversight/NomC 222.222366 |
        correct/Verb  oversight/NomC 139.799225 |
goof
         goof/NomC       goof/Verb 14860.539062 |
    egregious/AdjQ       goof/NomC 14447.746094 |
           mark/NomC       goof/NomC 556.276855 |
            big/AdjQ       goof/NomC 346.284180 |
blunder
       conceive/Verb    blunder/NomC 343.630310 |
[...]
```

Tableau 54 : *Coloc – liste L2-4 pour le nom 'erreur'*

4.2.1.3 Mise en correspondance des collocations en L1 et en L2

Il est maintenant plus facile d'établir des équivalences entre les deux listes de collocations (tableaux 50 et 54). Le tableau 55 met en correspondance des collocations en L1 (liste L1-4) et des traductions possibles (liste L2-4).

Équivalences possibles entre L1-4 et L2-4 pour le mot *erreur/NomC*	
Collocations L1	**Collocations équivalents L2**
erreur + monumental	grievous + error error + grievous grievous + mistake egregious + error error + egregious egregious + mistake egregious + goof
commettre + erreur	make + mistake error + execute
erreur + glisser	--
corriger + erreur rectifier + erreur	correct + error correct + flaw correct + oversight

Tableau 55 : *Coloc – équivalences possibles entre L1-4 et L2-4 pour le nom 'erreur'*

4.2.1.4 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 n'a pas d'équivalent. Qu'est-il advenu de collocations tout à fait habituelles comme *tomber + erreur* et *induire + erreur*? Pourquoi les équivalents *commit + mistake* ou *mistake + slip* n'ont-ils pas été extraits?

La raison est évidente. La liste utilisée pour le produit cartésien (et donc pour l'information mutuelle) comprenait (dans un cas comme dans l'autre) environ 170 mots. Ainsi, tout mot absent de cette liste n'était pas utilisé dans le calcul de l'information mutuelle. C'est le cas d'*induire*, *tomber*, *slip* et *commit*. Ces mots ne figurant pas sur la première liste et les calculs subséquents utilisant les éléments de cette liste, il est normal que ces mots n'apparaissent pas sur la liste de collocations.

Ce problème sera bientôt réglé. En effet, la version informatisée du modèle prendra tous les mots contenus dans les listes L1-1 et L2-1 (moins les mots exclus, bien entendu) pour effectuer le produit cartésien, et pas seulement les 100 ou 200 premiers mots. Davantage de collocations devraient donc être repérées en L1 et plusieurs choix de traductions extraits.

Bien entendu, les données extraites dépendent entièrement de la nature des textes eux-mêmes. Puisque le bitexte exploité ici est un corpus de nature plutôt orale[8], les collocations repérées ne sont pas aussi riches et nombreuses que si le bitexte avait été constitué de textes journalistiques ou littéraires par exemple. De plus, les collocations plus rares risquent de ne pas y apparaître. Idéalement, ces programmes devraient tourner sur un bitexte assez varié dans sa composition pour être représentatif de la langue prise dans son ensemble et assez grand pour se prêter facilement aux méthodes probabilistes.

---

[8] 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.

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

4.2.2 Étude de cas : *future/NomC*

L'exemple qui suit montre le traitement d'un autre substantif, en anglais cette fois. La procédure ne sera pas expliquée en détails comme pour *erreur/NomC*. Néanmoins, quelques précisions seront données pour faciliter la lecture des tableaux.

4.2.2.1 Extraction des collocations de future/NomC en L1

Le tableau 56 montre les 180 premiers résultats de *coloc* pour *future/NomC*.

Liste L1-1 pour *future/NomC*

future/NomC	0.679983	give/Verb	0.000955
future/AdjQ	0.051273	hold/Verb	0.000943
will/Verb	0.023141	allow/Verb	0.000937
country/NomC	0.014140	come/Verb	0.000918
development/NomC	0.011243	guarantee/Verb	0.000897
be/Verb	0.010775	just/Adve	0.000884
later/Adve	0.007307	so/Adve	0.000869
tomorrow/NomC	0.007124	near/AdjQ	0.000815
ensure/Verb	0.006501	people/NomC	0.000795
look/Verb	0.005193	nation/NomC	0.000792
build/Verb	0.004539	develop/Verb	0.000791
prepare/Verb	0.004537	past/NomC	0.000750
tomorrow/Adve	0.003991	promising/AdjQ	0.000750
day/NomC	0.003904	perspective/NomC	0.000700
now/Adve	0.003695	program/NomC	0.000692
vision/NomC	0.003262	sense/NomC	0.000657
forward/Adve	0.002732	good/AdjQ	0.000656
assure/Verb	0.002601	must/Verb	0.000650
opportunity/NomC	0.002472	still/Adve	0.000630
own/AdjQ	0.002354	accordingly/Adve	0.000594
prospect/NomC	0.002303	remain/Verb	0.000584
have/Verb	0.002302	not/Adve	0.000579
bright/AdjQ	0.002216	look forward to/Verb	0.000579
stake/NomC	0.002146	year/NomC	0.000575
ahead/Adve	0.002060	great/AdjQ	0.000562
see/Verb	0.001911	lie/NomC	0.000558
consider/Verb	0.001875	do/Verb	0.000553
face/Verb	0.001852	again/Adve	0.000535
road/NomC	0.001823	concern/Verb	0.000535
continue/Verb	0.001811	important/AdjQ	0.000514
make/Verb	0.001804	may/Verb	0.000506
look to/Verb	0.001709	think/Verb	0.000492
depend/Verb	0.001691	new/AdjQ	0.000487
provide/Verb	0.001687	way/NomC	0.000484
determine/Verb	0.001634	late/AdjQ	0.000482
sure/Adve	0.001629	take/Verb	0.000482
can/Verb	0.001518	also/Adve	0.000478
canadians/NomC	0.001513	kind/NomC	0.000458
generation/NomC	0.001490	concern/NomC	0.000454
canadian/AdjQ	0.001489	promise/NomC	0.000452
go/Verb	0.001420	as well/Adve	0.000450
common/AdjQ	0.001405	plan for/Verb	0.000438
be in/Verb	0.001303	assurance/NomC	0.000431
shape/Verb	0.001290	present/AdjQ	0.000419
secure/AdjQ	0.001278	commitment/NomC	0.000414
hope/NomC	0.001224	represent/Verb	0.000405
happen/Verb	0.001178	reserve/Verb	0.000401
stand/Verb	0.001123	plan/Verb	0.000401
reserve/NomC	0.001119	time/NomC	0.000389
secure/Verb	0.001085	ability/NomC	0.000385
shall/Verb	0.001073	expect/Verb	0.000384
become/Verb	0.001045	dependent/AdjQ	0.000372
today/Adve	0.001044	wait/Verb	0.000365
destiny/NomC	0.001006	far/AdjQ	0.000354
very/AdjQ	0.000987	threaten/Verb	0.000343

Liste L1-1 pour *future/NomC* (suite)			
plan/NomC	0.000338	keep/Verb	0.000214
believe/Verb	0.000337	foreseeable/AdjQ	0.000214
outlook/NomC	0.000333	decide/Verb	0.000205
hand/NomC	0.000328	possibility/NomC	0.000205
contemplate/Verb	0.000328	jeopardize/Verb	0.000204
direction/NomC	0.000319	compromise/Verb	0.000203
get/Verb	0.000307	show/Verb	0.000201
have in/Verb	0.000297	immediate/AdjQ	0.000198
able/AdjQ	0.000295	preparation/NomC	0.000198
well/Adve	0.000293	proposed/AdjQ	0.000194
reservation/NomC	0.000279	commit/Verb	0.000191
then/Adve	0.000272	livelihood/NomC	0.000190
hold/NomC	0.000272	deal/Verb	0.000190
world/NomC	0.000270	allot/Verb	0.000188
yet/Adve	0.000270	plight/NomC	0.000185
young/AdjQ	0.000269	certain/AdjQ	0.000183
long-term/AdjQ	0.000267	know/Verb	0.000181
fate/NomC	0.000265	do in/Verb	0.000179
look ahead/Verb	0.000264	come to/Verb	0.000178
promise/Verb	0.000261	uncertain/AdjQ	0.000176
turn/Verb	0.000259	rather/Adve	0.000175
involve/Verb	0.000257	mean/Verb	0.000174
want/Verb	0.000256	create/Verb	0.000160
affect/Verb	0.000249	help/Verb	0.000160
confidence/NomC	0.000248	mortgage/Verb	0.000159
very/Adve	0.000243	more/Adve	0.000153
consideration/NomC	0.000240	prospective/AdjQ	0.000152
let/Verb	0.000238	recently/Adve	0.000152
economic/AdjQ	0.000236	ready/AdjQ	0.000150
potential/AdjQ	0.000236	ago/Adve	0.000150
need/Verb	0.000232	date/NomC	0.000149
term/NomC	0.000230	want for/Verb	0.000147
developing/AdjQ	0.000229	policy/NomC	0.000144
potential/NomC	0.000220	issue/NomC	0.000143
here/Adve	0.000217	rely/Verb	0.000143
guarantee/NomC	0.000217		

Table 56 : ***Coloc -- liste L1-1 pour le nom 'future'***

Les mots contenus dans le tableau 57 ont été exclus de la liste précédente, afin de réduire le bruit dans les résultats produits à l'étape suivante.

Modifications apportées à L1-1 pour *future/NomC*	
Mots exclus (ordre alphabétique)	
all	much
also	must
as well	not
be	now
can	only
have	shall
just	very
may	well
more	will

Tableau 57 : *Coloc -- liste des mots exclus (L1-1) pour le nom 'future'*

De plus, les renseignements inutiles (comme la valeur de la probabilité) ont été supprimés.

Ainsi, en lançant *mtf_consult*, le modèle a effectué le produit cartésien de cette liste par elle-même, puis il a récupéré la valeur de l'information mutuelle pour chacun des couples ainsi jumelés. Les résultats suivants (tableau 58) représentent les 100 premières paires obtenues après le calcul de l'information mutuelle.

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

think/Verb	government/NomC	3604.736084
government/NomC	development/NomC	1391.808960
government/NomC	think/Verb	1294.454346
accordingly/Adve	tomorrow/NomC	719.554504
think/Verb	people/NomC	689.264465
development/NomC	government/NomC	547.157043
shape/Verb	destiny/NomC	546.677429
future/AdjQ	generation/NomC	494.008484
outlook/NomC	promising/AdjQ	369.018219
do/Verb	now/Adve	356.217377
foreseeable/AdjQ	future/NomC	348.541107
outlook/NomC	bright/AdjQ	343.677917
provide/Verb	people/NomC	318.753662
now/Adve	do/Verb	297.558197
mortgage/Verb	generation/NomC	283.767975
people/NomC	think/Verb	283.564148
think/Verb	make/Verb	246.429367
near/AdjQ	future/NomC	245.812241
canadian/AdjQ	country/NomC	244.624664
government/NomC	show/Verb	243.394913
government/NomC	year/NomC	236.848251
government/NomC	commitment/NomC	233.903687
contemplate/Verb	foreseeable/AdjQ	229.388306
look/Verb	government/NomC	220.688873
want/Verb	see/Verb	219.886108
stand/Verb	tomorrow/NomC	212.301651
uncertain/AdjQ	term/NomC	210.400467
tell/Verb	do/Verb	208.918823
prospective/AdjQ	foreseeable/AdjQ	208.871887
mortgage/Verb	future/NomC	208.219360
show/Verb	government/NomC	204.184677
today/Adve	canadian/AdjQ	198.808517
bright/AdjQ	outlook/NomC	196.387375
government/NomC	look/Verb	187.033813
people/NomC	provide/Verb	187.008987
bright/AdjQ	future/NomC	169.159607
common/AdjQ	sense/NomC	168.464783
year/NomC	government/NomC	166.320023
tomorrow/Adve	allot/Verb	164.510361
own/AdjQ	destiny/NomC	162.027863
depend/Verb	livelihood/NomC	155.107071
accordingly/Adve	stand/Verb	153.101837
allot/Verb	consideration/NomC	151.468155
prospect/NomC	foreseeable/AdjQ	150.960373
allot/Verb	day/NomC	147.999771
country/NomC	canadian/AdjQ	137.745605
do/Verb	want/Verb	137.459473
livelihood/NomC	depend/Verb	131.244446
government/NomC	good/AdjQ	130.515289
do/Verb	tell/Verb	130.064316

Liste L1-3 pour future/NomC (suite)		
government/NomC	go/Verb	127.760590
threaten/Verb	compromise/Verb	126.892303
go/Verb	get/Verb	126.176765
late/AdjQ	date/NomC	125.017319
promise/Verb	compromise/Verb	123.320282
make/Verb	provide/Verb	119.659172
government/NomC	economic/AdjQ	117.819397
commitment/NomC	government/NomC	116.168808
livelihood/NomC	stake/NomC	110.820152
mortgage/Verb	bright/AdjQ	108.714439
bright/AdjQ	mortgage/Verb	108.714439
shape/Verb	future/NomC	108.701149
bright/AdjQ	promising/AdjQ	108.013054
think/Verb	program/NomC	102.986412
bright/AdjQ	tomorrow/NomC	101.981873
government/NomC	program/NomC	100.260620
destiny/NomC	shape/Verb	99.395897
provide/Verb	make/Verb	97.840981
long/AdjQ	term/NomC	97.670723
government/NomC	great/AdjQ	96.963074
policy/NomC	people/NomC	95.499550
developing/AdjQ	world/NomC	94.771225
think/Verb	new/AdjQ	93.599327
think/Verb	give/Verb	91.959862
go/Verb	way/NomC	91.478065
foreseeable/AdjQ	developing/AdjQ	88.751427
developing/AdjQ	foreseeable/AdjQ	88.751427
want/Verb	do/Verb	88.056381
good/AdjQ	government/NomC	87.464005
uncertain/AdjQ	fate/NomC	86.255203
fate/NomC	uncertain/AdjQ	86.255203
great/AdjQ	government/NomC	84.729210
canadian/AdjQ	today/Adve	83.758385
generation/NomC	mortgage/Verb	83.461166
do/Verb	mean/Verb	83.188660
destiny/NomC	reserve/Verb	82.895393
determine/Verb	destiny/NomC	82.759094
people/NomC	policy/NomC	80.642365
government/NomC	young/AdjQ	80.310524
make/Verb	policy/NomC	80.009743
prospect/NomC	promising/AdjQ	78.402000
get/Verb	go/Verb	77.213760
mortgage/Verb	tomorrow/Adve	77.138840
do/Verb	get/Verb	76.811913
make/Verb	think/Verb	76.531548
decide/Verb	fate/NomC	75.979568
promising/AdjQ	outlook/NomC	73.803642
promising/AdjQ	future/NomC	71.831947
dependent/AdjQ	livelihood/NomC	72.452469
government/NomC	people/NomC	71.615471

Tableau 58 : *Coloc – liste L1-3 pour le nom 'future'*

Seuls les couples contenant le mot de départ, c'est-à-dire *future/NomC*, ont été retenus. Et il semble d'emblée que toutes ces combinaisons soient des collocations en bonne et due forme.

Liste L1-4 des combinaisons en L1 contenant le mot *future/NomC*

foreseeable/AdjQ	future/NomC	348.541107
near/AdjQ	future/NomC	245.812241
mortgage/Verb	future/NomC	208.219360
bright/AdjQ	future/NomC	169.159607
shape/Verb	future/NomC	108.701149
promising/AdjQ	future/NomC	71.831947
future/NomC	hold/NomC	70.388428
uncertain/AdjQ	future/NomC	62.039932
compromise/Verb	future/NomC	41.923656

Tableau 59 : *Coloc -- liste L1-4 pour le nom 'future'*

4.2.2.2 Extraction de collocations en L2 associées à future/NomC

Pour récupérer des collocations qui pourraient être équivalentes aux collocations trouvées à l'étape précédente, *future/NomC* a d'abord été donné en entrée à *matrix_consult*. Le tableau 60 contient les cent premiers équivalents de *future/NomC* trouvés dans le dictionnaire probabiliste.

Liste L2-1 pour *future/NomC*

#	203 : future/NomC	[NH	15k] #	0 : avenir/NomC	0.796260	#
#	203 : future/NomC	[NH	15k] #	1 : futur/AdjQ	0.035556	#
#	203 : future/NomC	[NH	15k] #	2 : assurer/Verb	0.015871	#
#	203 : future/NomC	[NH	15k] #	3 : développement/NomC	0.012380	#

```
#  203 : future/NomC    [NH   15k]#    4 : plus tard/Adve          0.011273 #
#  203 : future/NomC    [NH   15k]#    5 : pays/NomC               0.009212 #
#  203 : future/NomC    [NH   15k]#    6 : futur/NomC              0.007889 #
#  203 : future/NomC    [NH   15k]#    7 : préparer/Verb           0.007274 #
#  203 : future/NomC    [NH   15k]#    8 : demain/NomC             0.006136 #
#  203 : future/NomC    [NH   15k]#    9 : réserver/Verb           0.004783 #
#  203 : future/NomC    [NH   15k]#   10 : dorénavant/Adve         0.004445 #
#  203 : future/NomC    [NH   15k]#   11 : envisager/Verb          0.004321 #
#  203 : future/NomC    [NH   15k]#   12 : jour/NomC               0.003983 #
#  203 : future/NomC    [NH   15k]#   13 : perspective/NomC        0.003952 #
#  203 : future/NomC    [NH   15k]#   14 : prometteur/AdjQ         0.002907 #
#  203 : future/NomC    [NH   15k]#   15 : venir/Verb              0.002753 #
#  203 : future/NomC    [NH   15k]#   16 : désormais/Adve          0.002507 #
#  203 : future/NomC    [NH   15k]#   17 : dépendre/Verb           0.002414 #
#  203 : future/NomC    [NH   15k]#   18 : garantir/Verb           0.002291 #
#  203 : future/NomC    [NH   15k]#   19 : demain/Adve             0.002230 #
#  203 : future/NomC    [NH   15k]#   20 : réserve/NomC            0.002199 #
#  203 : future/NomC    [NH   15k]#   21 : continuer/Verb          0.002015 #
#  203 : future/NomC    [NH   15k]#   22 : être/Verb               0.001836 #
#  203 : future/NomC    [NH   15k]#   23 : sort/NomC               0.001813 #
#  203 : future/NomC    [NH   15k]#   24 : bâtir/Verb              0.001809 #
#  203 : future/NomC    [NH   15k]#   25 : génération/NomC         0.001782 #
#  203 : future/NomC    [NH   15k]#   26 : encore/Adve             0.001705 #
#  203 : future/NomC    [NH   15k]#   27 : en jeu/Adve             0.001363 #
#  203 : future/NomC    [NH   15k]#   28 : résider/Verb            0.001267 #
#  203 : future/NomC    [NH   15k]#   29 : canadien/AdjQ           0.001240 #
#  203 : future/NomC    [NH   15k]#   30 : entrevoir/Verb          0.001217 #
#  203 : future/NomC    [NH   15k]#   31 : compromettre/Verb       0.001113 #
#  203 : future/NomC    [NH   15k]#   32 : pouvoir/Verb            0.001048 #
#  203 : future/NomC    [NH   15k]#   33 : tourner/Verb            0.001009 #
#  203 : future/NomC    [NH   15k]#   34 : attendre/Verb           0.000978 #
#  203 : future/NomC    [NH   15k]#   35 : devenir/Verb            0.000963 #
#  203 : future/NomC    [NH   15k]#   36 : devenir/NomC            0.000925 #
#  203 : future/NomC    [NH   15k]#   37 : concerner/Verb          0.000905 #
#  203 : future/NomC    [NH   15k]#   38 : songer/Verb             0.000832 #
#  203 : future/NomC    [NH   15k]#   39 : devoir/Verb             0.000763 #
#  203 : future/NomC    [NH   15k]#   40 : programme/NomC          0.000740 #
#  203 : future/NomC    [NH   15k]#   41 : menacer/Verb            0.000613 #
#  203 : future/NomC    [NH   15k]#   42 : propre/AdjQ             0.000598 #
#  203 : future/NomC    [NH   15k]#   43 : faire/Verb              0.000582 #
#  203 : future/NomC    [NH   15k]#   44 : main/NomC               0.000563 #
#  203 : future/NomC    [NH   15k]#   45 : préoccuper/Verb         0.000544 #
#  203 : future/NomC    [NH   15k]#   46 : représenter/Verb        0.000529 #
#  203 : future/NomC    [NH   15k]#   47 : possibilité/NomC        0.000525 #
#  203 : future/NomC    [NH   15k]#   48 : aussi/Adve              0.000521 #
#  203 : future/NomC    [NH   15k]#   49 : hypothéquer/Verb        0.000509 #
#  203 : future/NomC    [NH   15k]#   50 : constituer/Verb         0.000498 #
#  203 : future/NomC    [NH   15k]#   51 : prendre/Verb            0.000486 #
#  203 : future/NomC    [NH   15k]#   52 : permettre/Verb          0.000475 #
#  203 : future/NomC    [NH   15k]#   53 : relève/NomC             0.000459 #
#  203 : future/NomC    [NH   15k]#   54 : avoir/Verb              0.000459 #
```

```
#  203 : future/NomC    [NH    15k]#   55 : reproduire/Verb          0.000429 #
#  203 : future/NomC    [NH    15k]#   56 : un jour ou l'autre/Adve 0.000429 #
#  203 : future/NomC    [NH    15k]#   57 : falloir|faillir/Verb    0.000425 #
#  203 : future/NomC    [NH    15k]#   58 : ne/Adve                 0.000402 #
#  203 : future/NomC    [NH    15k]#   59 : immédiat/NomC           0.000398 #
#  203 : future/NomC    [NH    15k]#   60 : lendemain/NomC          0.000398 #
#  203 : future/NomC    [NH    15k]#   61 : destinée/NomC           0.000398 #
#  203 : future/NomC    [NH    15k]#   62 : canadiens/NomC          0.000394 #
#  203 : future/NomC    [NH    15k]#   63 : brillant/AdjQ           0.000386 #
#  203 : future/NomC    [NH    15k]#   64 : à long terme/Adve       0.000386 #
#  203 : future/NomC    [NH    15k]#   65 : engager/Verb            0.000371 #
#  203 : future/NomC    [NH    15k]#   66 : décider/Verb            0.000363 #
#  203 : future/NomC    [NH    15k]#   67 : présent/AdjQ            0.000356 #
#  203 : future/NomC    [NH    15k]#   68 : offrir/Verb             0.000348 #
#  203 : future/NomC    [NH    15k]#   69 : davantage/Adve          0.000340 #
#  203 : future/NomC    [NH    15k]#   70 : plus/Adve               0.000336 #
#  203 : future/NomC    [NH    15k]#   71 : année/NomC              0.000336 #
#  203 : future/NomC    [NH    15k]#   72 : voie/NomC               0.000329 #
#  203 : future/NomC    [NH    15k]#   73 : meilleur/AdjQ           0.000317 #
#  203 : future/NomC    [NH    15k]#   74 : occuper/Verb            0.000310 #
#  203 : future/NomC    [NH    15k]#   75 : promettre/Verb          0.000279 #
#  203 : future/NomC    [NH    15k]#   76 : plan/NomC               0.000275 #
#  203 : future/NomC    [NH    15k]#   77 : pas moins/Adve          0.000263 #
#  203 : future/NomC    [NH    15k]#   78 : éventuellement/Adve     0.000256 #
#  203 : future/NomC    [NH    15k]#   79 : déterminant/AdjQ        0.000235 #
#  203 : future/NomC    [NH    15k]#   80 : influer/Verb            0.000229 #
#  203 : future/NomC    [NH    15k]#   81 : orientation/NomC        0.000226 #
#  203 : future/NomC    [NH    15k]#   82 : actuel/AdjQ             0.000222 #
#  203 : future/NomC    [NH    15k]#   83 : donner/Verb             0.000220 #
#  203 : future/NomC    [NH    15k]#   84 : confiance/NomC          0.000215 #
#  203 : future/NomC    [NH    15k]#   85 : plutôt/Adve             0.000214 #
#  203 : future/NomC    [NH    15k]#   86 : promesse/NomC           0.000210 #
#  203 : future/NomC    [NH    15k]#   87 : aller/Verb              0.000209 #
#  203 : future/NomC    [NH    15k]#   88 : aujourd'hui/Adve        0.000191 #
#  203 : future/NomC    [NH    15k]#   89 : projeter/Verb           0.000186 #
#  203 : future/NomC    [NH    15k]#   90 : utile/AdjQ              0.000185 #
#  203 : future/NomC    [NH    15k]#   91 : rapproché/AdjQ          0.000183 #
#  203 : future/NomC    [NH    15k]#   92 : vouloir/Verb            0.000179 #
#  203 : future/NomC    [NH    15k]#   93 : nouveau/AdjQ            0.000174 #
#  203 : future/NomC    [NH    15k]#   94 : immédiat/AdjQ           0.000172 #
#  203 : future/NomC    [NH    15k]#   95 : espoir/NomC             0.000167 #
#  203 : future/NomC    [NH    15k]#   96 : ensemble/Adve           0.000166 #
#  203 : future/NomC    [NH    15k]#   97 : important/AdjQ          0.000164 #
#  203 : future/NomC    [NH    15k]#   98 : à nouveau/Adve          0.000160 #
#  203 : future/NomC    [NH    15k]#   99 : grand/AdjQ              0.000160 #
#  [...]                                                                    #
#=================================#=======================================#
```

Tableau 60 : *Coloc – liste L2-1 pour le nom 'future'*

Ensuite, certains mots ont été manuellement éliminés de cette liste (voir tableau 61); seuls les mot_i/cat_gram_i ont été gardés.

Mots exclus de la liste L2-1 pour *future/NomC* avant le calcul de l'information mutuelle	
ainsi	pas
avoir	plus
bien	si
ci	tant
être	tout
ne	très

Tableau 61 : ***Coloc – liste des mots exclus (L2-1) pour le nom 'future'***

À l'étape suivante, *mtf_consult* effectue d'abord le produit cartésien de cette dernière liste par elle-même, puis associe une valeur d'information mutuelle à chacune des paires, comme le montre le tableau 61.

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

devoir/Verb	besoin/NomC	3510.141357
besoin/NomC	devoir/Verb	1533.103149
gouvernement/NomC	mesure/NomC	1311.900635
gouvernement/NomC	mettre/Verb	1286.209473
savoir/Verb	nouveau/AdjQ	946.546448
hypothéquer/Verb	futur/NomC	870.171265
mesure/NomC	gouvernement/NomC	865.073181
génération/NomC	futur/AdjQ	606.966797
brillant/AdjQ	postérité/NomC	510.309082
mettre/Verb	gouvernement/NomC	471.030365
nouveau/AdjQ	pays/NomC	420.369720
aller/Verb	tenir/Verb	404.533875
avenir/NomC	rapproché/AdjQ	390.276672
lointain/AdjQ	devenir/NomC	379.638580
tenir/Verb	aller/Verb	368.697815
décider/Verb	faire/Verb	341.773499
affronter/Verb	postérité/NomC	327.869690
immédiat/NomC	destinée/NomC	319.347778
devoir/Verb	faire/Verb	315.242920
savoir/Verb	faire/Verb	314.913605
nouveau/AdjQ	savoir/Verb	305.807343
prendre/Verb	mesure/NomC	275.878052
hypothéquer/Verb	avenir/NomC	263.214539
hypothéquer/Verb	génération/NomC	256.389740
pays/NomC	nouveau/AdjQ	248.163803
faire/Verb	programme/NomC	221.284271
faire/Verb	pays/NomC	220.881897
immédiat/NomC	futur/NomC	211.260834
canadiens/NomC	pouvoir/Verb	208.969833
pouvoir/Verb	canadiens/NomC	205.082657
mesure/NomC	prendre/Verb	193.614304
augurer/Verb	incertain/AdjQ	191.551483
faire/Verb	savoir/Verb	179.745651
futur/NomC	prometteur/AdjQ	176.759613
tâcher/Verb	devenir/NomC	175.293564
futur/NomC	hypothéquer/Verb	174.034241
avenir/NomC	prometteur/AdjQ	172.511993
faire/Verb	intérêt/NomC	167.515533
immédiat/NomC	rapproché/AdjQ	167.462860
programme/NomC	faire/Verb	163.217728
ultérieur/AdjQ	rapproché/AdjQ	162.537476
gouvernement/NomC	pouvoir/Verb	161.950790
boucher/Verb	éventuellement/Adve	157.731903
faire/Verb	décider/Verb	157.466324
futur/NomC	destinée/NomC	153.123169
avenir/NomC	incertain/AdjQ	144.720734
augurer/Verb	avenir/NomC	144.238190
entrevoir/Verb	brillant/AdjQ	142.827042
immédiat/NomC	incertain/AdjQ	134.188477
augurer/Verb	prometteur/AdjQ	131.557739

Liste L2-3 pour *future/NomC* (suite)

futur/NomC	incertain/AdjQ	128.683304
voir/Verb	canadien/AdjQ	127.311127
nouveau/AdjQ	canadiens/NomC	127.293030
pays/NomC	faire/Verb	127.246666
pouvoir/Verb	faire/Verb	125.837425
sombre/AdjQ	lointain/AdjQ	124.722763
pouvoir/Verb	gouvernement/NomC	122.908707
faire/Verb	devoir/Verb	121.701233
immédiat/NomC	entrevoir/Verb	120.104546
survie/NomC	dépendre/Verb	119.522202
bâtir/Verb	ensemble/Adve	118.779381
résolument/Adve	tourner/Verb	118.431946
tourner/Verb	résolument/Adve	118.431946
voir/Verb	gouvernement/NomC	116.241447
relève/NomC	demain/NomC	114.404053
devoir/Verb	pouvoir/Verb	114.081406
optique/NomC	augurer/Verb	113.965706
avérer/Verb	utile/AdjQ	113.714493
clé/NomC	résider/Verb	113.584473
résolument/Adve	affronter/Verb	109.289894
faire/Verb	nécessaire/AdjQ	109.037560
vouloir/Verb	maintenant/Adve	107.305573
jour/NomC	lendemain/NomC	105.242477
perspective/NomC	sombre/AdjQ	104.878357
déterminant/AdjQ	rapproché/AdjQ	104.038383
hypothéquer/Verb	futur/AdjQ	103.817169
sort/NomC	réserver/Verb	103.739532
bâtir/Verb	prometteur/AdjQ	102.917854
affronter/Verb	incertain/AdjQ	102.526024
gouvernement/NomC	décider/Verb	102.313469
sort/NomC	incertain/AdjQ	101.145363
faire/Verb	projet/NomC	100.336098
prospérité/NomC	futur/AdjQ	99.603409
attendre/Verb	gouvernement/NomC	97.947304
canadiens/NomC	nouveau/AdjQ	97.830338
avenir/NomC	lointain/AdjQ	97.545845
stable/AdjQ	viable/AdjQ	96.500389
intérêt/NomC	faire/Verb	95.614967
lointain/AdjQ	bâtir/Verb	93.982185
génération/NomC	hypothéquer/Verb	93.232635
projet/NomC	faire/Verb	92.201492
faire/Verb	canadiens/NomC	91.400352
perspective/NomC	prometteur/AdjQ	91.185516
postérité/NomC	main/NomC	90.012451
déterminant/AdjQ	relève/NomC	88.558624
gouvernement/NomC	offrir/Verb	87.139587
aussi/Adve	pays/NomC	86.761787
faire/Verb	genre/NomC	85.969147
vouloir/Verb	prendre/Verb	85.185272
futur/NomC	immédiat/AdjQ	85.019600

Tableau 62 : *Coloc – liste L2-3 pour le nom 'future'*

4.2.2.3 Mise en correspondance des collocations en L1 et L2

Les paires ont été regroupées autour des substantifs qui pourraient être des équivalents français de *future/NomC* tableau 63). Ainsi, les couples contenant *avenir* ont été placés ensemble, suivi des paires comprenant *futur*, et ainsi de suite.

Liste L2-4 pour *future/NomC*
regroupée autour par substantif (extrait)

```
avenir
      avenir/NomC  rapproché/AdjQ 390.276672 |
 hypothéquer/Verb     avenir/NomC 263.214539 |
      avenir/NomC prometteur/AdjQ 172.511993 |
      avenir/NomC  incertain/AdjQ 144.720734 |
     augurer/Verb     avenir/NomC 144.238190 |
       avenir/NomC   lointain/AdjQ 97.545845 |
    entrevoir/Verb     avenir/NomC 71.465591 |
  perspective/NomC     avenir/NomC 63.690399 |
        bâtir/Verb     avenir/NomC 60.161686 |
      avenir/NomC hypothéquer/Verb 53.992729 |
futur
 hypothéquer/Verb      futur/NomC 870.171265 |
       futur/NomC prometteur/AdjQ 176.759613 |
      futur/NomC hypothéquer/Verb 174.034241 |
       futur/NomC   destinée/NomC 153.123169 |
       futur/NomC  incertain/AdjQ 128.683304 |
        futur/NomC   immédiat/AdjQ 85.019600 |
      optique/NomC      futur/NomC 76.561584 |
      tourner/Verb      futur/NomC 56.222641 |
perspective
 perspective/NomC     sombre/AdjQ 104.878357 |
  perspective/NomC prometteur/AdjQ 91.185516 |
  perspective/NomC   brillant/AdjQ 64.801155 |
sort
        sort/NomC   réserver/Verb 103.739532 |
        sort/NomC  incertain/AdjQ 101.145363 |
   préoccuper/Verb       sort/NomC 75.149338 |
génération
  génération/NomC      futur/AdjQ 606.966797 |
 hypothéquer/Verb génération/NomC 256.389740 |
  génération/NomC hypothéquer/Verb 93.232635 |
       futur/AdjQ génération/NomC 83.626534 |
destinée
    immédiat/NomC   destinée/NomC 319.347778 |
  éventualité/NomC   destinée/NomC 84.826744 |
survie
      survie/NomC   dépendre/Verb 119.522202 |
      assurer/Verb     survie/NomC 71.948853 |
       survie/NomC    menacer/Verb 63.889191 |
    essentiel/AdjQ     survie/NomC 63.839359 |
 compromettre/Verb     survie/NomC 59.881981 |
[...]
```

Tableau 63 : *Coloc – liste L2-4 pour le nom 'future'*

Lorsque les deux listes L1-4 et L2-4 sont mises en correspondance, les résultats sont très intéressants pour le lexicographe bilingue (tableau 64).

Équivalences possibles entre L1-4 et L2-4 pour *future/NomC*	
Collocations L1	**Collocations équivalents L2**
foreseeable + future near + future	avenir + rapproché futur + immédiat immédiat + destinée
mortgage + future compromise + future	hypothéquer + avenir avenir + hypothéquer hypothéquer + futur futur + hypothéquer hypothéquer + génération génération + hypothéquer survie + menacer compromettre + survie
bright + future promising + future	avenir + prometteur futur + prometteur perspective + prometteur perspective + brillant
shape + future	bâtir + avenir assurer + survie
future + hold	augurer + avenir sort + réserver
uncertain + future	avenir + incertain futur + incertain sort + incertain perspective + sombre

Tableau 64 : ***Coloc – équivalences possibles entre L1-4 et L2-4 pour le nom 'future'***

4.2.2.4 Critique

Comme pour *erreur/NomC*, la liste de collocations en L1 n'est pas exhaustive. Cependant, tel que nous l'avons expliqué dans la section 4.2.1.4, ce problème sera réglé, en partie à tout le moins, lorsque le modèle aura été informatisé, puisque les listes ne seront pas tronquées à quelques centaines de mots.

Par contre, il est intéressant de constater que toutes les collocations en L1 ont au moins 2 équivalents, ce qui offre au lexicographe quelques options.

Les résultats, ici comme dans le cas de *erreur/NomC*, sont fonction du corpus lui-même et, comme l'atteste la fréquence de *future/NomC* (plus de 15 000), on parle beaucoup d'avenir à la Chambre des communes. Lorsque le modèle informatisé sera au point, il sera intéressant d'évaluer les résultats produits par ce modèle dans le cas de mots moins fréquents.

4.3 Informatisation du modèle

Nous travaillons présentement à l'intégration de toutes les étapes décrites à la section 4.2 en un modèle complet. Quoique la programmation ne soit pas encore tout à fait au point, notre prototype donne déjà des résultats encourageants, et tout porte à croire que cet outil, qui extraira, pour un *mot/cat_gram* donné, une série de collocations en L1 ainsi que des traductions possibles en L2, constituera un ajout de taille au poste de travail du lexicographe bilingue.

Pour l'instant, le prototype fonctionne séparément pour la L1 et la L2. Ainsi, le lexicographe lance deux requêtes, une première visant à produire une liste de collocations possibles en L1[9] et l'autre, une liste d'équivalents possibles de ces collocations en L2[10].

La génération automatique de la liste des collocations potentielles en L1 suit les mêmes étapes que la méthodologie expliquée à la section 4.2. Après que le lexicographe a tapé *mot/cat_gram langue*, cette information est donnée en entrée à *coloc*, qui produit une liste de mots compilée à partir des deux dictionnaires bilingues probabilistes[11]. De cette liste ne sont conservés que les mot_i/cat_gram_i qui n'apparaissent pas dans le fichier des mots exclus[12], sans aucun autre élément d'information. *Mtf_consult* prend ensuite cette liste de mots, fait le produit cartésien de cette liste par elle-même et cherche, dans une table d'information mutuelle déjà calculée, la valeur de l'information mutuelle pour toutes les combinaisons trouvées. Finalement, seuls les couples contenant le mot de départ *mot/cat_gram* sont retenus, et ils sont affichés en ordre décroissant d'information mutuelle.

Voici les résultats obtenus automatiquement avec la commande *liste1.sh* pour les deux mots que nous avons étudiés ici, soit *erreur/NomC f* (tableau 65) et *future/NomC e* (tableau 66).

---

[9] Avec la commande suivante « liste1.sh *mot/cat_gram langue* » (*liste1.sh erreur/NomC f* et *liste1.sh future/NomC e* pour les cas que nous avons déjà analysés).

[10] Avec la commande « liste2.sh *mot/cat_gram langue* » (*liste2.sh erreur/NomC f* et *liste2.sh future/NomC e* dans les cas qui nous intéressent ici).

[11] Ces dictionnaires ont été décrits à la sous-section 4.1.1.

[12] Il est à noter que ce fichier peut être enrichi *ad hoc*.

liste1.sh erreur/NomC f

induire/Verb	erreur/NomC	1070.3635
erreur/NomC	monumental/AdjQ	271.4440
commettre/Verb	erreur/NomC	178.1452
erreur/NomC	glisser/Verb	160.1406
erreur/NomC	inadvertance/NomC	155.3843
corriger/Verb	erreur/NomC	118.0847
erreur/NomC	commettre/Verb	115.0323
inadvertance/NomC	erreur/NomC	110.4046
rectifier/Verb	erreur/NomC	98.4776
grossier/AdjQ	erreur/NomC	95.6840
erreur/NomC	omission/NomC	86.9600
erreur/NomC	passé/AdjQ	76.2509
bévue/NomC	erreur/NomC	68.5907
erreur/NomC	bévue/NomC	68.5907

Tableau 65 : ***Modèle automatisé – liste 1 pour le nom 'erreur'***

liste1.sh future/NomC e

foreseeable/AdjQ	future/NomC	348.5411
near/AdjQ	future/NomC	245.8122
mortgage/Verb	future/NomC	208.2194
bright/AdjQ	future/NomC	169.1596
shape/Verb	future/NomC	108.7011
promising/AdjQ	future/NomC	71.8319
uncertain/AdjQ	future/NomC	62.0399
prosperous/AdjQ	future/NomC	53.2617
brilliant/AdjQ	future/NomC	45.1813
vision/NomC	future/NomC	44.1138

Tableau 66 : ***Modèle automatisé – liste 1 pour le nom 'future'***

Il est évident que ni l'une ni l'autre de ces listes n'est exhaustive. Mais là n'est pas la question. L'important, ici, c'est que ces listes aient été générées de façon tout à fait automatique, sans aucune intervention du lexicographe[13].

Le prototype affiche aussi une série de collocations possibles en L2 avec la commande *liste2.sh mot/cat_gram langue.* D'abord, le prototype donne *mot/cat_gram* en entrée à *matrix_consult* qui, lui, repère dans la matrice L1-->L2 tous les équivalents possibles en L2. De cette liste d'équivalents, seuls les mot_i/cat_gram_i sont conservés (la probabilité associée à chacun de ces mots n'est pas conservée pour l'étape suivante). En outre, tous les mots énumérés dans le fichier de mots exclus[14] sont éliminés de la liste. Ensuite, *mtf_consult* fait le produit cartésien de cette liste par elle-même et cherche dans une table d'information mutuelle pré-calculée la valeur de l'information mutuelle pour toutes les combinaisons ainsi créées. Les deux tableaux suivants constituent les 25 premières paires fournies par ce programme pour les deux mots à l'étude ici, soit *erreur/NomC f* (tableau 67) et *future/NomC e* (tableau 68).

---

[13] Dans le cadre de cette thèse, nous n'évaluerons pas la rendement de ce modèle en termes de *précision* et de *rappel*. La *précision* est la mesure de la qualité des éléments repérés (nombre d'éléments valables divisé par le nombre d'éléments repérés), et le *rappel* est la mesure de l'efficacité du système (nombre d'éléments valables divisé par le nombre total d'éléments valables). Cette étude fera bientôt l'objet d'une communication conjointe avec le CITI.

[14] Le fichier des mots exclus peut être modifié au fur et à mesure des besoins.

liste2.sh erreur/NomC f *(extrait)*

goof/NomC	goof/Verb	14860.5391
egregious/AdjQ	goof/NomC	14447.7461
factually/Adve	incorrect/AdjQ	3709.4570
egregious/AdjQ	error/NomC	2970.6528
factually/Adve	untrue/Verb	1690.5272
sin/NomC	omission/NomC	1601.7582
mistake/Verb	untrue/NomC	1215.9886
grievous/AdjQ	error/NomC	1208.4011
brunt/NomC	miscalculation/NomC	963.6590
house/NomC	question/NomC	926.6707
mistakenly/Adve	execute/Verb	915.0578
quasi-judicial/AdjQ	body/NomC	895.5617
clerical/AdjQ	error/NomC	894.1114
factual/AdjQ	erroneously/Adve	821.2403
sin/NomC	time-consuming/AdjQ	758.9283
question/NomC	house/NomC	701.6753
inadvertently/Adve	mislead/Verb	701.1339
factually/Adve	correct/Verb	678.1210
gather/Verb	glitch/NomC	651.2883
fallacy/NomC	misconception/NomC	603.1063
drafting/NomC	oversight/NomC	561.3717
mark/NomC	goof/NomC	556.2769
misconception/NomC	misunderstanding/NomC	545.7415
erroneously/Adve	execute/Verb	505.6898
detect/Verb	wrong/NomC	482.7986
[...]		

Tableau 67 : *Modèle automatisé – liste 2 pour le nom 'erreur'*

liste2.sh future/NomC e *(extrait)*

finance/NomC	dire/Verb	6909.4541
planche/NomC	salut/NomC	5402.9956
dire/Verb	finance/NomC	3552.4116
garant/NomC	futur/NomC	1053.4873
hypothéquer/Verb	futur/NomC	870.1713
génération/NomC	futur/AdjQ	606.9668
brillant/AdjQ	postérité/NomC	510.3091
revêtir/Verb	importance/NomC	459.1367
demander/Verb	accord/NomC	427.9697
nouveau/AdjQ	pays/NomC	420.3697
avenir/NomC	rapproché/AdjQ	390.2767
lointain/AdjQ	devenir/NomC	379.6386
affronter/Verb	postérité/NomC	327.8697
immédiat/NomC	destinée/NomC	319.3478
ministre/NomC	environnement/NomC	316.8135
peu/Adve	importer/Verb	316.6046
édification/NomC	hypothéquer/Verb	303.9106
passé/NomC	garant/NomC	289.3381
brillant/AdjQ	carrière/NomC	283.1305
faire/Verb	domaine/NomC	276.9696
prendre/Verb	mesure/NomC	275.8781
hypothéquer/Verb	avenir/NomC	263.2145
ministre/NomC	penser/Verb	256.6179
hypothéquer/Verb	génération/NomC	256.3897
pays/NomC	nouveau/AdjQ	248.1638
[...]		

Tableau 68 : ***Modèle automatisé – liste 2 pour le nom 'future'***

Il est bien évident que ces listes sont difficiles à utiliser, car les équivalents potentiels y sont présentés en vrac. Ainsi, nous préparons présentement une petite « moulinette » qui regroupera ces paires selon la fréquence du composant le plus fréquent. Le modèle trouvera d'abord le mot qui revient le plus souvent dans la liste et affichera les couples qui contiennent ce mot en ordre décroissant d'information mutuelle. Il cherchera ensuite le mot qui revient le plus souvent dans les paires qui restent et affichera, à la suite des paires déjà trouvées, les paires qui contiennent ce

nouveau mot. La « moulinette » répétera cette opération jusqu'à ce qu'il ne reste plus aucune paire dans la liste (les paires dont les composants seront uniques se retrouveront ensemble, à la fin de la liste triée). Cette étape est présentement effectuée manuellement, mais voici comment les résultats se présenteront lorsque cette « moulinette » sera au point.

```
liste2.sh mot/cat_gram langue

mot1/cat_de_mot1
  couples contenant mot1
mot2/cat_de_mot2
  couples contenant mot2
mot3/cat_de_mot3
  couples contenant mot3
[...]
```

Tableau 69 : *Modèle automatisé – présentation optimale*

Bien évidemment, les listes générées en L1 et en L2 à partir d'un mot en L1 sont intimement liées puisque la liste en L1, de par la façon dont elle est calculée (avec *coloc*), est tributaire de la liste de collocations en L2. Ainsi, des associations sémantiques sont possibles entre les deux.

Le diagramme présenté à la page suivante (tableau 70) illustre la procédure utilisée dans le cadre du projet d'informatisation du modèle.

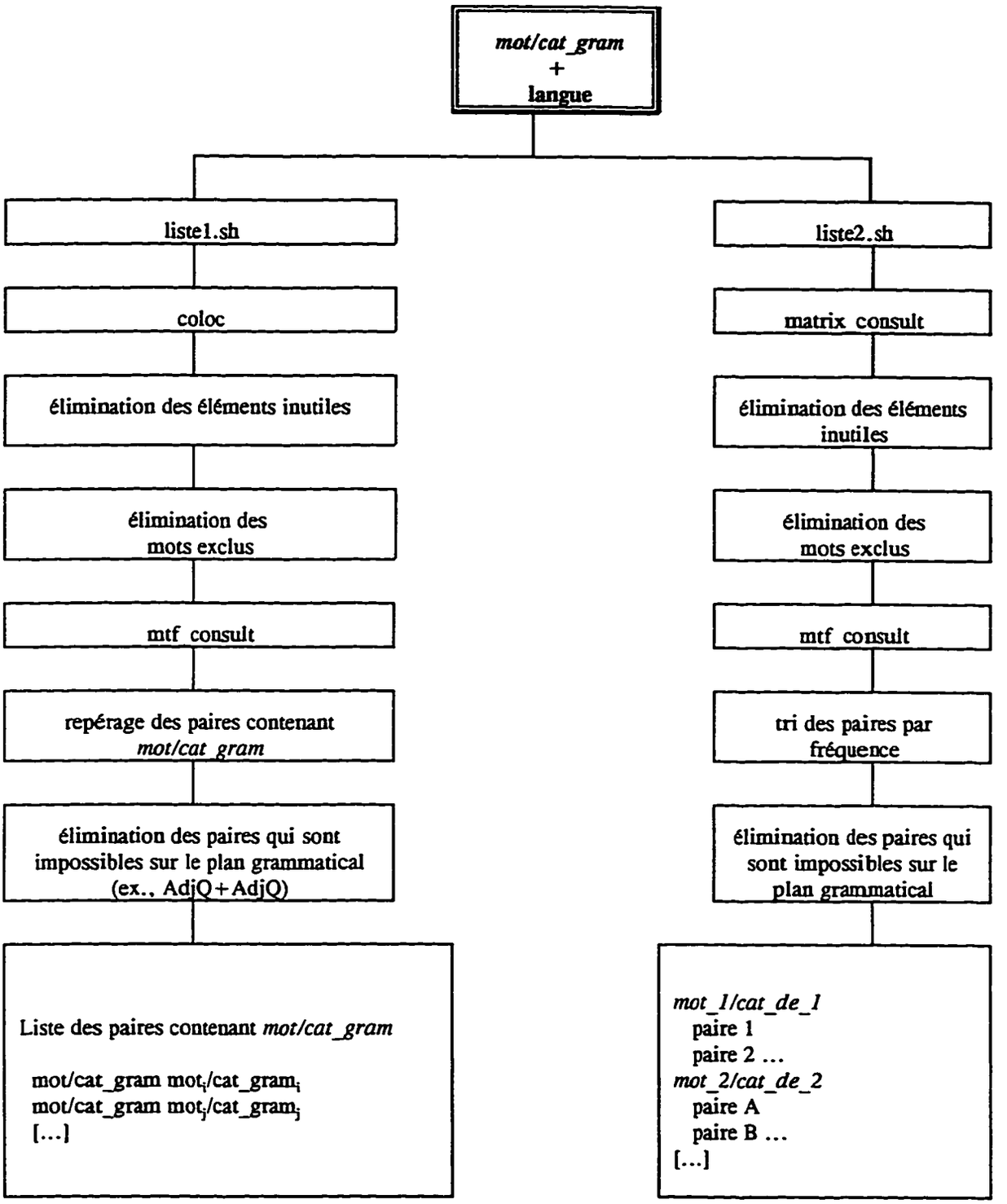

Tableau 70 : *Modèle automatisé*

Nous en sommes encore au stade des essais. Nous tentons d'optimiser le rendement du modèle en changeant différents paramètres, à savoir, par exemple, combien de mots et/ou de couples devraient être retenus après chaque calcul, la meilleure façon d'effectuer les calculs ou encore quels mots devraient figurer sur la liste de mots exclus.

Nous sommes plutôt satisfaits des résultats pour la Liste L1. Malheureusement, la liste en L2 est plus problématique. Nous cherchons un moyen heuristique de réduire la quantité de collocations potentielles que cette liste contient. Nous avons pensé à modifier un peu la façon dont cette liste est calculée. En ce moment, le modèle cherche, pour un mot qui, disons, serait un substantif (*mot/NomC*), tous les équivalents dans le dictionnaire bilingue probabiliste; ensuite le modèle élimine les mots contenus dans la liste des mots exclus, puis fait le produit cartésien de cette liste par elle-même. Qu'arriverait-il si, au lieu de ce produit cartésien, on faisait plutôt le produit cartésien de cette liste par la liste des substantifs contenus sur cette liste et non sur la liste complète? Puisque, avec notre modèle, nous n'extrayons que des paires contenant le *mot/NomC* en L1, il va de soi que les équivalents susceptibles d'intéresser le lexicographe bilingue contiendront aussi un $mot_i/NomC_i$.

4.4 Comparaison de notre modèle avec Champollion

Il est difficile de comparer notre modèle à celui de Smadja. D'une part, les deux programmes ne tournent pas sur le même bitexte et, d'autre part, nos recherches n'ont pas été effectuées avec les mêmes mots. Cela dit, nous sommes d'avis que la principale raison pour laquelle nous ne

pouvons comparer directement nos résultats à ceux de Smadja est liée au fait que son modèle et le nôtre ne produisent pas le même type d'information : le modèle de Smadja ne prend en entrée qu'un mot ou une collocation et produit le mot ou la collocation qui est plus susceptible d'en être la traduction, tandis que notre modèle donne, pour un mot, toute une série de collocations possibles en L1 et toute une série de collocations qui pourraient en être les traductions.

Champollion serait sans doute plus utile en traduction qu'en lexicographie, puisqu'il permet à l'utilisateur de chercher une traduction en L2 étant donné un mot ou une collocation en L1. Ainsi, un traducteur qui n'arrive pas à penser d'emblée à la meilleure façon de traduire une collocation trouverait sans doute Champollion utile puisque ce programme permet de résoudre des problèmes de traduction au cas par cas.

En revanche, si le lexicographe ne travaille que sur un seul mot à la fois, il en analyse néanmoins toutes les combinatoires possibles. Ainsi, notre modèle lui convient-il davantage puisqu'il propose une série de collocations en L1 ainsi que toute une série de traductions potentielles en L2. Ce qui est intéressant aussi, au niveau des traductions potentielles, c'est que cette liste renferme des collocations synonymiques (par exemple, pour *to launch an attack*, on trouvera *prendre l'offensive, lancer une attaque, prendre d'assaut*) qu'il serait impossible d'extraire autrement. Il est évident que le lexicographe doit choisir la meilleure façon de présenter toutes ces possibilités dans son entrée, mais, les possibilités sont bien là, il n'a pas à les chercher.

Notre modèle, cependant, présente le désavantage de se limiter à des paires de mots. Or, si les collocations apparaissent généralement sous cette forme, il est aussi possible que l'équivalent d'une collocation soit constitué d'un seul mot. Ainsi, pour la collocation *launch/Verb attack/NomC*, notre modèle trouve facilement *passer/Verb attaque/NomC* et *prendre/Verb offensive/NomC*, mais il ne donnera jamais *attaquer/Verb* comme équivalent, bien que ce verbe apparaisse sur la liste L2-1 produite par *matrix_consult*. Ainsi, à moyen terme, nous aimerions trouver un moyen d'inclure des équivalents *simples* à la liste d'équivalents possibles en L2.

CONCLUSION

Dans le cadre de ce travail, nous nous étions fixé comme premier objectif de montrer dans quelle mesure les bitextes pouvaient être utiles en lexicographie bilingue. Nous avons d'abord suggéré diverses façons dont les bi-concordanciers, et les bi-concordances qu'ils produisent, peuvent servir la cause du lexicographe bilingue, notamment pour accroître le nombre d'équivalents présentés dans une entrée, pour confirmer un équivalent proposé dans un dictionnaire, pour traduire une collocation ou une expression figée. De plus, puisque notre bitexte est un corpus exclusivement canadien, le lexicographe peut l'utiliser pour repérer des canadianismes.

Notre deuxième objectif était d'élaborer un logiciel pour l'extraction de collocations et de leurs traductions possibles. Ainsi, nous avons conçu un programme qui produit automatiquement deux listes de collocations potentielles, une en L1, où L1 est la langue du mot demandé, et l'autre en L2. La première liste en L1 n'affiche que des collocations possibles contenant le mot demandé en entrée; la liste en L2, elle, est constituée de toute une panoplie de combinaisons de mots qui pourraient être des équivalents des combinaisons contenues dans la liste en L1.

Même si les bi-concordanciers et l'extracteur de collocations donnent des résultats intéressants, il n'en demeure pas moins qu'ils ne sont, et ne seront jamais, que des outils. Aussi le lexicographe doit-il se servir de son jugement lorsqu'il les utilise. Dans le cas des bi-concordanciers, par exemple, il peut arriver que certains contextes soient beaucoup trop spécifiques pour pouvoir être utilisés d'emblée dans une entrée ou encore qu'une paire de phrases

repérée par le logiciel contienne une erreur flagrante de traduction (par exemple, nous avons déjà vu le substantif anglais *physician* traduit par *physicien* en français).

Cette mise en garde vaut aussi pour l'extracteur de collocations. Si les listes en L1 ne contiennent que très peu de bruit et sont donc faciles à utiliser, elles sont, en revanche, loin d'être exhaustives. Ainsi, le lexicographe qui utilise cet outil devra sans doute, pour l'instant à tout le moins, enrichir l'entrée avec des collocations qu'il aura obtenues ailleurs. Les listes en L2, par contre, contiennent énormément de bruit, et le lexicographe doit les utiliser avec soin, choisir manuellement les meilleures traductions et éliminer celles qui sont superflues.

Cela dit, il est évident que ces deux outils ouvrent des possibilités extraordinaires pour l'avenir. Il est facile d'imaginer, par exemple, l'intégration d'un bi-concordancier comme TransSearch et d'un bitexte au poste de travail du traducteur. Le traducteur pourrait les utiliser de façon ponctuelle pour l'aider à résoudre des problèmes de traduction particuliers. En outre, le bitexte *Hansard* pourrait servir de mémoire de traduction s'il était utilisé avec un logiciel de traduction comme *TranslationManager* d'IBM.

L'extracteur de collocations ouvre aussi la porte à diverses possibilités en lexicographie et en traduction automatique. Grâce à lui, il serait facile de compiler un dictionnaire bilingue voué exclusivement aux collocations, par exemple. Aucun autre modèle, à notre connaissance, ne peut donner des séries synonymiques potentielles en L2 (et en L1, si nous conservons toutes les paires trouvées et pas seulement celles qui contiennent le mot de départ), et c'est souvent de ces

possibilités que les rédacteurs et traducteurs ont besoin lorsqu'ils rédigent ou traduisent. De plus, il est évident que les dictionnaires de transfert des systèmes de traduction automatique seraient grandement enrichis par l'apport d'un dictionnaire bilingue de collocations.

Nous croyons avoir montré que les bitextes et les outils qui ont été conçus spécialement pour les exploiter sont d'une aide précieuse pour le lexicographe bilingue. Si le bi-concordancier TransSearch en est arrivé à une version *commercialisable*, ce n'est pas encore le cas pour notre extracteur de collocations. Malgré cela, nous avons montré qu'un extracteur de collocations, même à l'étape du prototype, donne des résultats supérieurs à l'information extraite par un concordancier et, du coup, facilite énormément le travail du lexicographe.

GLOSSAIRE[1]

bi-concordance
: extrait d'un bitexte qui contient des segments dans une langue accompagnés de leur traduction

bi-concordancier
: logiciel d'analyse textuelle conçu pour produire des bi-concordances

bitexte
: corpus dans lequel les segments équivalents d'un texte de départ et de sa traduction sont mis en correspondance

collocation
: combinaison de deux ou plusieurs mots dans laquelle les composants, même s'ils conservent leur autonomie sémantique, sont soumis à des contraintes lexicales

concordance
: extrait d'un corpus dans lequel le mot qui fait l'objet de l'analyse est placé dans son contexte naturel

condordancier
: logiciel d'analyse textuelle conçu pour produire des concordances

corpus
: ensemble de textes informatisés, parlés ou écrits, qui servent de point de départ pour la description linguistique ou pour prouver une hypothèse sur la langue

corpus bilingue
: voir *corpus parallèle*

corpus comparable
: corpus constitué de textes écrits dans deux langues différentes qui, sans être des traductions mutuelles, fonctionnent de façon semblable au plan de la situation de communication

corpus de traduction
: corpus constitué de textes de départ et de leurs traductions

corpus parallèle
: corpus constitué de textes de départ et de leurs traductions ou corpus constitué de textes écrits dans deux langues différentes qui, sans être des traductions mutuelles, fonctionnent de façon semblable au plan de la situation de communication (synonyme : *corpus bilingue*)

extracteur de collocations
: logiciel qui sert à repérer des collocations dans un corpus

---

[1] Certaines définitions sont originales, et d'autres sont des adaptations de définitions proposées par d'autres chercheurs.

BIBLIOGRAPHIE

Ouvrages et articles cités

AARTS, J. (1990): « Corpus Linguistics: An Appraisal », *Computers in Literary and Linguistic Research - Proceedings of the Fifteenth International Conference*, Y. Choueka (Éd.), VILLE, Champion-Slatkine, pp. 13-27.

ALLERTON, D.J. (1984): « Three (or Four) Levels of Word Cooccurrence Restriction », *Lingua*, W. S. Allen, J.G. Kooij, E.C. Garcia (Éds.), Amsterdam, Elsevier Science Publishers, vol. 63, pp. 17-40.

AISENSTADT, E. (1979): « Collocability Restrictions in Dictionaries », *Review of Applied Linguistics*, vol. 45-46, pp. 71-74.

ARNOLD, D., BALKAN, L., LEE HUMPHREYS, R., MEIJER, S., et S. SADLER (1994): *Machine Translation - An Introductory Guide*, Manchester/Oxford, NCC Blackwell.

BARNBROOK, G. (1996): *Language and Computers*, Édimbourg, Edinburgh University Press.

BAUGH, S., HARLEY, A. et S. JELLIS (1996): « Compiling the Cambridge Dictionary of English », *International Journal of Corpus Linguistics*, vol. 1, n° 1, pp. 39-60.

BENDER, T.K. (1977): « Innovations in the Format of Literary Concordances and Indexes », *Linguistics - An International Review*, La Haye/Paris/New York, Mouton, vol. 197, pp. 53-63.

BENSON, P. (1994): « Electronic text analysis and the lexis of Hong Kong English: what can it tell us that we don't already know? », *Entering Text*, L. Flowerdew & A.K.K. Tong (Éds.), Hong Kong, Language Centre, The Hong Kong University of Science and Technology, 1994, pp. 89-101.

BIBER, D. (1994): « Using Register-Diversified Corpora for General Language Studies », *Using Large Corpora,* S. Armstrong (Éd.), Cambridge, MIT Press, pp. 179-202.

BROWN, P.F., DELLA PIETRA, S.A., DELLA PIETRA, V.J. et R.L. MERCER (1993): « The Mathematics of Statistical Machine Translation: Parameter Estimation », *Computational Linguistics*, vol. 19, n° 2, pp. 263-311.

CHURCH, K.W. et W.A. GALE (1991a): « A Program for Aligning Sentences in Bilingual Corpora », *COLING 91 - Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 177-184.

CHURCH, K.W. et W.A. GALE (1991b): « Concordances for Parallel Text », *Proceedings of the 7th Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, Oxford, pp. 1-14.

CHURCH, K.W. et HANKS, P. (1990): « Word Association Norms, Mutual Information and Lexicography », *Computational Linguistics*, vol. 16, n° 1, pp. 22-29.

CLAS, A. (1994): « Collocations et langues de spécialité », *Meta*, vol. 39, n° 4, pp. 576-580.

CLEAR, J. (1986a): « Trawling the language: Monitor Corpora », *ZuriLEX'86 Proceedings - Papers read at the EURALEX International Congress*, M. Snell-Hornby (Éd.), Francke Verlag, pp. 383-389.

CLEAR, J.(1986b): « Computers, Corpora and Modern Lexicography: the COBUILD Experience », *Computers in Literary and Linguistic Research - Proceedings of the Fifteenth International Conference*, Y. Choueka (Éd.), Champion-Slatkine, 1990, pp. 93-105.

CLEAR, J. (1987): « Overview of the Role of Computing in Cobuild », *Looking Up - An Account of the COBUILD Project in Lexical Computing*, J.M. Sinclair (Éd.), Londres, HarperCollins, pp. 41-61.

CLEAR, J. (1990): « Computers, Corpora and Modern Lexicography », *Computers in Literary and Linguistic Research - Proceedings of the Fifteenth International Conference*, Y. Choueka (Éd.), Paris/Genève, Champion-Slatkine, pp. 93-105.

COP, M. (1988): « The Function of Collocations in Dictionaries », *BudaLEX '88 Proceedings -Papers from the Third International EURALEX Congress*, T. Magay et J. Zigany (Éds.), Budapest, pp. 35-46.

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

DAUPHIN, É. (1994): « Étude de corpus : un préalable pour l'adaptation des systèmes de traduction automatique aux besoins des utilisateurs », *TA-TAO : Recherches de pointe et applications immédiates*, A. Clas et P. Bouillon (Éd.), Montréal, AUPELF-UREF, pp. 17-25.

DESCAMPS, J.-L. (1994): « Tournoi pour l'accommodement des dictionnaires de collocations », *Meta*, vol. 39, n° 4, pp. 561-575.

FIRTH, J.R. (1951): « Modes of Meaning », *Papers in Linguistics*, F.R. Palmer (Éd.), Londres, Oxford University Press, pp. 190-215.

FONTENELLE, T. (1994): « What on earth are collocations? », *English Today*. vol. 10, n° 4, pp. 42-48.

FRANCIS, N.W. (1967): « The Brown University Standard Corpus of English: Some Implications for TESOL », *On Teaching English to Speakers of Other Languages*, B.W. Robinett (Éd.), Washington, Teachers of English to Speakers of Other Languages, pp. 131-137.

FRANCIS, N.W. (1992): « Language Corpora B.C. », *Directions in Corpus Linguistics - Proceedings of Nobel Symposium 82*, J. Svartvik (Éd.), Berlin/New York, Mouton de Gruyter, pp. 17-31.

HANON, S. (1990a): « La concordance », *Dictionaries: An International Encyclopaedia of Lexicography*, F.J Hausmann *et al.* (Éds.), Berlin/New York, Walter de Gruyter, pp. 1562-1566.

HANON, S. (1990b): *Le Vocabulaire de l'« Heptaméron » de Marguerite de Navarre*, Paris-Genève, Champion-Slatkine, 1990.

HARRIS, B. (1988): « Bi-text, a new concept in translation theory », *Language Monthly*, n° 54, pp. 8-10.

HARTMANN, R.R.K. (1980): *Contrastive Textology. Comparative Discourse Analysis in Applied Linguistics (Studies in Descriptive Linguistics 5)*, J. Gross (Éd.), Heidelberg.

HAUSMANN, F.J. (1990): « Le dictionnaire de collocations », *Dictionaries: An International Encyclopaedia of Lexicography*, F.J Hausmann *et al.* (Éds.), Berlin/New York, Walter de Gruyter, pp. 1010-1019.

HUTCHINS, W.J. et SOMERS, H.L. (1992): *An Introduction to Machine Translation*, Londres, Academic Press.

ISABELLE, P. et S. WARWICK-ARMSTRONG (1993): « Les corpus bilingues : une nouvelle ressource pour le traducteur », *La Traductique*, P. Bouillon et A. Clas (Éds.), Montréal, Les Presses de l'Université de Montréal, pp. 288-308.

LANGLOIS, L. (1996): « Bilingual Concordancers: A New Tool for Bilingual Lexicographers », *Expanding MT Horizons - Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, pp. 34-42.

LEECH, G. (1991): « The State of the Art in Corpus Linguistics », *English Corpus Linguistics*, K. Aijmer et B. Altenberg (Éds.), Londres/New York, Longman, pp. 8-29.

LEECH, G. (1992): « Corpora and Theories of Linguistic Performance », *Directions in Corpus Linguistics - Proceedings of Nobel Symposium 82*, J. Svartvik (Éd.), Berlin/New York, Mouton de Gruyter, pp. 105-120.

LEECH, G. et S. FLIGELSTONE (1992): « Computers and Corpus Analysis », *Computers and the Written Text*, C.S. Butler (Éd.), Oxford/Cambridge, Blackwell, pp. 115-139.

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

LIPSHITZ, E. (1981): « La Nature sémanto-structurelle des phraséologismes analytiques verbaux », *Cahiers de lexicologie*, vol. 38, n° 1, pp. 35-44.

MARCUS, M. P., ASANTORINI, B. et M.A. MARCINKIEWICZ (1994): «Building a Large Annotated Corpus of English: The Penn Treebank », *Using Large Corpora*, S. Armstrong (Éd.), Cambridge, MIT Press, pp. 273-290.

McENERY, T. et A. WILSON (1996): *Corpus Linguistics*, Édimbourg, Edinburgh University Press.

MEIJS, W. (1982): « Exploring Brown with 'QUERY' », *Computer Corpora in English Language Research*, S. Johansson (Éd.), Bergen, Norwegian Computing Centre for the Humanities, pp. 34-38.

PERGNIER, M. (1980): *Les fondements sociolinguistiques de la traduction*, Paris, H. Champion.

RENOUF, A. (1987): « Corpus Development », *Looking UP - An Account of the COBUILD Project in Lexical Computing*, J.M. Sinclair (Éd.), Londres, HarperCollins, pp. 1-40.

ROBERTS, R.P. (1994/1995): « Identifying the Phraseology of Languages for Special Purposes », *Actes de langue française et de linguistique*, vol. 7/8, pp. 61-74.

ROBERTS, R.P. (1996): « Parallel-Text Analysis and Bilingual Lexicography » (à paraître).

RUNDELL, M. et P. STOCK (1992a): « The Corpus Revolution (Part 1) », *English Today*, n° 30, pp. 9-14.

RUNDELL, M. et P. STOCK (1992b): « The Corpus Revolution (Part 2) », *English Today*, n° 31, pp. 21-32.

RUNDELL, M. et P. STOCK (1992c): « The Corpus Revolution (Part 3) », *English Today*, vol. 32, pp. 45-51.

RUNDELL, M. (1995): « The Word on the Street », *English Today*, n° 43, p. 29-35.

SAUSSURE, F. de (1964): *Cours de linguistique générale*, Paris, Payot.

SIMARD, M., FOSTER, G. et P. ISABELLE (1992): « Using Cognates to Align Sentences in Bilingual Corpora », *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 67-81.

SINCLAIR, J.M. (1982): « Reflections on Computer Corpora in English Language Research », *Computer Corpora in English Language Research*, S. Johansson (Éd.), Bergen, Norwegian Computing Centre for the Humanities, pp. 1-6.

SINCLAIR, J.M. (1987a) (Éd.): *Looking Up - An Account of the COBUILD Project in Lexical Computing*, Londres, HarperCollins.

SINCLAIR, J.M. (1987b): « Collocation: a progress report », *Laguage Topics - Essays in Honour of Michael Halliday - Vol. II*, R. Steele et T. Threadgood (Éds.), Amsterdam/ Philadelphie, John Benjamins, pp. 319-331.

SINCLAIR, J.M. (1991): *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

SINCLAIR, J.M. (1992): « The Automatic Analysis of Corpora », *Directions in Corpus Linguistics - Proceedings of Nobel Symposium 82*, J. Svartvik (Éd.), Berlin/New York, Mouton de Gruyter, pp. 379-400.

SLABY, W.A. (1978): « Concordances to the Greek New Testatement and to the 'Bad' Quartos of the works of Shakespeare: two strategies for an automatic selection of context ». *Advances in Computer-Aided Literary and Linguistic Research, Proceedings of the Fifth International Symposium on Computers in Literary and Linguistic Research*, D.E. Ager, F.E. Knowles et J. Smith (Éds.), pp. 117-127.

SMADJA, F. (1993): « Retrieving Collocations from Text: Xtract », *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 », *Computational Linguistics*, vol. 22, n° 1, pp. 1-38.

SVARTVIK, J. (1992): « Corpus Linguistics comes of age », *Directions in Corpus Linguistics - Proceedings of Nobel Symposium 82*, Jan Svartvik (Éd.), Berlin/New York, Mouton de Gruyter, pp. 7-13.

VINAY, J.-P. et DARBELNET, J. (1958): *Stylistique comparée du français et de l'anglais*, Beauchemin.

Ouvrages et articles consultés

ATKINS, B.T.S. et B. Levin: « Building on a Corpus: A Linguistic and Lexicographical Look at some Near-Synonyms », *International Journal of Lexicography*, vol. 8, n° 2, 1995, pp. 85-144.

BENSON, M.: « Collocations and Idioms » *Dictionaries, Lexicography and Language Learning*, R. Ilson (Éd.), Oxford, Pergamon Press, 1985, pp. 61-68.

CLEAR, J.: « From Firth Principles - Computational Tools for the Study of Collocation », *Text and Technology - In Honour of John Sinclair*, Philadelphie/Amsterdam, John Benjamins, 1993, pp. 271-292.

COP, M.: « Collocations in the Bilingual Dictionary », *Dictionaries: An International Encyclopaedia of Lexicography*, F.J Hausmann *et al.* (Éd.), Berlin/New York, Walter de Gruyter, pp. 2775-2778.

COWIE, A.P.: « The Treatment of Collocations and Idioms in Learners' Dictionaries », *Applied Linguistics*, vol. 2-3, 1981, pp. 223-235.

COWIE, A.P. et P. HOWARTH.: « Phraseology - A Select Bibliography », *International Journal of Lexicography*, vol. 9, n° 2, pp. 118-131.

FONTENELLE, T.: « Towards the Construction of a Collocational Database of Translation Students », *Méta*, vol. 39, n° 1, pp. 47-56.

FRANCIS, W.N.: « A Tagged Corpus - Problems and Prospects », *Studies in English Linguistics for Randolph Quirk*, S. Greenbaum, G. Leech et J. Svartvik (Éds.), Londres/New York, Longman, 1980, pp. 192-209.

FRANCIS, W.N.: « Problems of Assembling and Computerizing Large Corpora », *Computer Corpora in English Language Research*, S. Johansson (Éd.), Bergen, Norwegian Computing Centre for the Humanities, 1982, pp. 7-24.

GAATONE, D.: « La locution ou le poids de la diachronie dans la synchronie », *La locution - Actes du colloque international*, Montréal, Université McGill, 1984, pp. 71-81.

ISABELLE, P.: « La bi-textualité : vers une nouvelle génération d'aides à la traduction et la terminologie », *Meta*, vol. 37, n° 4, 1992, pp. 721-739.

ISABELLE, P.: « Bi-Textual Aids for Translators », rapport technique, CITI (1575, boul. Chomedey, Laval, Québec, H7V 2X2).

KJELLMER, G.: « Some Problems Relating to the Study of Collocations in the Brown Corpus », *Computer Corpora in English Language Research*, S. Johansson (Éd.), Bergen, Norwegian Computing Centre for the Humanities, 1982, pp. 25-33.

KJELLMER, G.: « Some Thoughts on Collocational Distinctiveness », *Corpus Linguistics - Recent Developments in the Use of Computer Corpora in English Language Research*, J. Aarts et W. Meijs (Éds.), Amsterdam, Rodopi, 1984, pp. 163-172.

LANGLOIS, L.: *Rapport sur l'utilisation de TransSearch en lexicographie bilingue*, rapport technique, Université d'Ottawa, 1995.

LEONARD, R.: « The Computer Archive of Modern English Texts », *Computational and Mathematical Linguistics - Proceedings of the International Conference on Computational Linguistics*, Firenze, L.S. Olschki (Éd.), 1977, pp. 417-428.

MACKIN, R.: « On Collocations : 'Words shall be known by the company the keep' », *In Honour of A.S. Hornby*, P. Strevens (Éd.), Oxford, Oxford University Press, pp. 149-165.

MACKLOVITCH, E.: « Corpus-based Tools for Translators », *Proceedings of the 33rd Annual Conference of the American Translators Association*, 1992, pp. 317-328.

MEL'CUK, I.: *Dictionnaire explicatif et combinatoire du français contemporain*, Montréal, Les Presses de l'Université de Montréal, 1984.

QUIRK, R.: « On Corpus Principles and Design », *Directions in Corpus Linguistics - Proceedings of Nobel Symposium 82*, J. Svartvik (Éd.), Berlin/New York, Mouton de Gruyter, pp. 457-469.

SIMARD, M., FOSTER, G.F. et F. PERRAULT: *TransSearch: A Bilingual Concordance Tool*, rapport technique, CITI (1575, boul. Chomedey, Laval, Québec, H7V 2X2).

VAN SCHERRENBURG, D.: *The Arrangement of Information in the General Bilingual Dictionary Entry*. Thèse de maîtrise, Université d'Ottawa, 1990.

Dictionnaires consultés

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

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

Dictionnaire Hachette-Oxford français-anglais/anglais-français, Oxford, Oxford University Press, 1994.

French-English/English French Unabridged Dictionary, Paris, Larousse, 1993.

LACROIX, U. (1956): *Dictionnaire des mots et des idées*, Paris, Fernand Nathan.

MARTINET, A. (Éd.) (1973): *Le langage*, Paris, Les Dictionnaires du Savoir Moderne, 1973.

MOUNIN, G. (1974): *Dictionnaire de la linguistique*, Paris, Quadrige/Presses Universitaires de France.

Nouveau Petit Robert - Dictionnaire de la langue française, Paris, Dictionnaires Le Robert, 1993.

PEI, M. et GAYNOR (1954): *A Dictionary of Linguistics*, New York, Philosophical Library.

PHELIZON, J.F. (1976): *Vocabulaire de la linguistique*, Paris, Éditions Roudil.

Random House Webster's College Dictionary, New York, Random House, 1991.

Robert & Collins Dictionnaire français-anglais anglais-français Senior, 3e édition, Paris, Dictionnaires Le Robert, 1993.