Université de Montréal

Automatisation du repérage et de l'encodage des collocations en langue de spécialité

par
Brigitte Orliac
Département de linguistique et de traduction
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en linguistique

juillet 2004

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

Automatisation du repérage et de l'encodage
des collocations en langue de spécialité

présentée par :

Brigitte Orliac

a été évaluée par un jury composé des personnes suivantes :

Richard Kittredge, président-rapporteur
Alain Polguère, directeur de recherche
Marie-Claude L'Homme, co-directeur
Patrick Drouin, membre du jury
Béatrice Daille, examinateur externe
Guy Lapalme, représentant du doyen de la FES

Thèse acceptée le : _____

Résumé

Les travaux présentés dans ces pages visent le développement d'une méthode d'extraction des collocations verbales de la langue de spécialité. Elle s'appuie sur un modèle de description formel des collocations, celui des fonctions lexicales de la théorie Sens-Texte.

Afin d'extraire les collocatifs verbaux des termes, nous avons analysé un corpus informatique anglais de 600 000 occurrences à l'aide du système de traduction automatique Logos. L'analyse du corpus a produit des arbres syntaxiques dont le niveau le plus élevé représente les principaux constituants de la phrase : le verbe et ses dépendants syntaxiques.

Nous avons développé des règles qui identifient les structures syntaxiques de surface caractéristiques des collocations verbales et extraient les combinaisons formées du verbe et de l'un de ses dépendants des arbres produits par l'analyseur de Logos. Les règles d'extraction sont regroupées dans trois grammaires différentes, pour chacune des trois relations syntaxiques modélisées par les fonctions lexicales verbales. Elles sont intégrées dans un programme (Colex) qui calcule également la fréquence de chaque combinaison extraite. L'extraction syntaxique des combinaisons verbales donne de bons résultats puisque 85 % environ des combinaisons extraites représentent une des trois relations syntaxiques majeures.

Finalement nous avons mis en place un filtre statistique afin d'isoler les collocations véritables des combinaisons extraites sur des bases

syntaxiques. Nous avons évalué la précision de deux mesures d'association sur les combinaisons extraites par Colex. Les critères sémantiques retenus pour l'évaluation étant assez contraignants, la précision du meilleur filtre atteint 71 %. Nous pensons qu'un filtre mesurant l'aptitude des combinaisons verbales à s'encoder dans des fonctions lexicales standard permettrait d'obtenir de meilleurs résultats.

Mots-clés : collocations, fonctions lexicales, extraction de collocations, statistique lexicale, langue de spécialité, terminologie.

Abstract

The work presented here focuses on the development of a semi-automatic method for extracting specialized verb + noun collocations. It is based on lexical functions, the formal device developed within the framework of the Meaning-Text theory to represent collocations.

In order to extract the verbal collocates of terms, we analyzed a specialized corpus of 600 000 occurrences. The corpus was analyzed by the Logos Machine Translation system. Analysis by the translation system produced syntactic trees whose higher nodes represent the major sentence constituents: the verb and its syntactic dependents.

We developed rules which identify the surface syntactic structures of verbal collocations and extract the combinations formed by the verb and a dependent from the trees produced by Logos. The extraction rules are organized in three grammars, one for each of the three major syntactic relations. They are implemented in a program (Colex) which also computes the frequency of each extracted combination. Precision of the syntactically-based extraction of verb + noun combinations is high (85 %).

Finally, we implemented a statistical filter to isolate the true collocations in the lists of syntactically-extracted combinations. We evaluated the performance of two association measures in selecting the collocations among the combinations extracted by Colex. Semantic criteria used in evaluating the combinations explain the precision of the better measure (71 %). We believe that a filter based on measuring a

combination's ability to be encoded as a lexical function would achieve higher precision.

Keywords : collocations, lexical functions, collocation extraction, lexical statistics, specialized language, terminology.

Remerciements

Je tiens à remercier mon directeur, Alain Polguère, à qui je dois d'avoir entrepris et mené jusqu'au bout ce travail. Plus que tout, son amitié et sa confiance m'ont permis de surmonter (**Real**) cette redoutable épreuve.

Je tiens aussi à remercier Marie-Claude L'Homme qui a joué un rôle beaucoup plus important que celui de co-directrice. Marie a suivi mes progrès pas à pas, me relevant à chaque fois et me remettant sur la voie. À elle aussi, je dois ce travail.

Je remercie également les membres du jury, et particulièrement Patrick Drouin et Béatrice Daille, pour tous les commentaires qu'ils ont faits, qui m'ont indiqué de nombreuses pistes vers lesquelles orienter le travail présenté dans ces pages.

Un immense merci à Mike Dillinger et à Esmé Manandise qui ont chacun programmé des parties importantes de Colex. Mike a permis que l'outil voit le jour, en obtenant l'autorisation de Logos et en programmant les idées que j'avais développées sur papier (et toutes les révisions que je lui ai envoyées). Esmé a développé en une semaine un outil de traitement statistique des données sur la base de deux longues conversations téléphoniques.

Finalement, je veux remercier ma famille et tous les amis qui ont su me redonner courage lorsque je voulais tout envoyer promener. Je leur dois de pouvoir enfin écrire ces lignes.

À Christian Orliac

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

BBI.....*BBI Combinatory Dictionary of English*

BNC.....British National Corpus

COBUILD.....*Collins COBUILD Advanced Learner's English Dictionary*

DAFA.....*Dictionnaire d'apprentissage du français des affaires*

DAFLES*Dictionnaire d'apprentissage du français langue étrangère ou seconde*

DEC.....Dictionnaire explicatif et combinatoire

DiCo.....*dictionnaire combinatoire*

FL.....fonction lexicale

LAF.....*Lexique actif du français*

LDOCE.....*Longman Dictionary of Contemporary English*

LEC.....lexicologie explicative et combinatoire

OALDCE.....*Oxford Advanced Learner's Dictionary of Current English*

OLST.....Observatoire de linguistique Sens-Texte

RES.....*Resolution module*

SAL.....*Semantico-syntactic Abstraction Language*

TA.....traduction automatique

TAL.....traitement automatique de la langue

TRAN.....*Transfer module*

TST.....théorie Sens-Texte

USS.....unité sémantico-syntaxique

1 Introduction

Cette recherche a pour but la construction d'un modèle d'acquisition semi-automatique de connaissances lexicales en vue d'applications en traitement automatique de la langue (TAL). Dans cette perspective, nous nous intéressons particulièrement au repérage et à l'encodage des collocations, ces expressions d'un caractère conventionnel, particulièrement fréquentes dans la langue. Ainsi, des expressions usitées telles que *fierce battle*, *strong resistance*, *excruciating pain* mais aussi *[to] pay attention*, *[to] spark an interest* ou *[to] strike a deal* sont toutes des exemples de collocations de l'anglais.

En développant un modèle d'extraction de collocations, nous voulons également contribuer à l'étude de la langue de spécialité dont les modes de combinaison privilégiés sont encore relativement méconnus.

1.1 Importance des phénomènes collocationnels

Les collocations, ou expressions semi-idiomatiques, appartiennent, avec les expressions idiomatiques, à l'ensemble des combinaisons lexicales non libres d'une langue. Comme les expressions idiomatiques, elles associent les lexies à l'intérieur des phrases, selon des schémas pré-établis qui doivent être appris par les locuteurs. Elles se distinguent des expressions entièrement idiomatiques par l'interprétation sémantique que l'on peut en faire. Alors que le sens d'une expression idiomatique est largement incompréhensible pour celui qui entend l'expression pour la première fois (il ne peut jamais être dérivé du sens des unités lexicales qui composent l'expression), le sens d'une collocation est généralement

accessible, bien que de façon moins directe que celui des énoncés composés librement : comparer, par exemple, les collocations *fierce battle* ou *[to] pay attention* données ci-dessus à des expressions entièrement idiomatiques telles que *[to] crack somebody up* ('faire rire qqn'), *[to] tip the scales* ('avoir une influence décisive') ou *[to] get one's head together* ('s'efforcer à considérer logiquement une situation déconcertante'). La transparence sémantique des collocations est due en grande partie à l'unité lexicale qui garde son sens à l'intérieur de la combinaison, unité choisie librement par le locuteur et que l'on considère pour cette raison comme base de la collocation. Le sens « problématique » est celui exprimé par le collocatif, dans l'entourage seul de la base, sens qui est particulier à l'expression elle-même et pour cette raison partiellement obscur (cf. le sens de *fierce* ou de *pay* dans les expressions *fierce battle* ou *[to] pay attention*).

Plus encore que le caractère non compositionnel de son interprétation sémantique, c'est le caractère contraint de la sélection de ses constituants qui définit véritablement une collocation. Ainsi, l'expression transparente *[to] brush one's teeth* est une collocation en anglais. Pour exprimer la même action, on ne peut employer une expression équivalente telle que *[to] wash one's teeth* (les deux expressions sont possibles en français)[1].

Les collocations abondent dans la langue. Les chercheurs, informaticiens et linguistes, qui s'intéressent à leur description, constatent qu'elles sont en général dix fois plus nombreuses, dans le lexique d'une

---

[1] Nous empruntons ces exemples à Mel'čuk (2003).

langue donnée, que les unités lexicales simples (Gross 1982; Chanier *et al.* 1995; Mel'čuk 1998). Leur connaissance est primordiale pour un système de traduction automatique (TA). En effet, à cause de leur sémantisme particulier, elles ne peuvent être traduites littéralement, sur la base seule des lexies qui les composent. Le système de TA qui ignore le sens de l'adjectif *fierce* ou du verbe *pay* dans les expressions *fierce battle* ou *[to] pay attention* traduit selon le sens que ces lexies ont habituellement et produit les expressions non idiomatiques suivantes (la traduction correcte est indiquée entre parenthèses):

(1) *fierce battle* ⇒ **bataille féroce (bataille acharnée)*

(2) *[to] pay attention* ⇒ **payer l'attention (faire attention)*

1.2 Collocations dans le discours spécialisé

Les traductions non idiomatiques reproduites ci-dessus sont encore plus nombreuses lorsque le système de traduction automatique est utilisé pour la traduction de textes techniques. En effet, bien qu'ils soient surtout appelés à traduire des textes techniques, les systèmes de TA sont généralement mal adaptés à la langue de spécialité. L'adaptation, lorsqu'elle est faite, consiste à créer des dictionnaires spécialisés contenant une majorité d'entrées nominales. Nous donnons quelques exemples pris au hasard dans un dictionnaire d'informatique :

(3) *character data*
cyberspace
gigabyte
host

IP address
multimedia
prompt
random access
unicode
webcast
etc.

Or, le discours spécialisé ne se caractérise pas seulement par des termes, unités nominales simples ou complexes, mais aussi par des constructions linguistiques plus larges qui, en structurant le discours spécialisé, facilitent elles aussi la communication entre experts :

(4) *[to] call a program*
[to] issue a command
[to] kill a process
[to] save a file
[to] store data on a disk
etc.

Ces modes de combinaison privilégiés du discours spécialisé sont rarement représentés dans les systèmes de TA. Les verbes sont particulièrement mal traités puisque, à moins d'avoir fait l'objet d'une entrée particulière dans le dictionnaire spécialisé du système (le domaine servant alors à délimiter entièrement le sens du verbe spécialisé), ils seront traduits selon le sens qu'ils ont en langue générale :

(5) ***Cache memory stores*** *both instructions and data* ⇒ **La mémoire cache emmagasine les instructions aussi bien que*

les données (La mémoire cache contient les instructions aussi bien que les données)

(6) *You can* ***enter*** *and* ***run a request*** ⇒ **Vous pouvez introduire et diriger une demande (Vous pouvez entrer et exécuter une demande)*

Les combinaisons lexicales en gras dans les exemples ci-dessus (qui illustrent des relations prédicatives entre un verbe et ses arguments) sont, au même titre que les termes d'un domaine spécialisé, les moyens d'expression conventionnels de ce domaine, moyens mis en œuvre par les différents spécialistes du domaine pour communiquer. Il s'agit d'expressions linguistiques non libres qui, telles les collocations de la langue générale, doivent être répertoriées dans les dictionnaires des systèmes de TA.

Alors que des ouvrages documentent exclusivement les collocations de la langue générale (cf. le *BBI Combinatory Dictionary of English* (*BBI*) (Benson *et al.* 1997) ou le *Collins COBUILD Advanced Learner's English Dictionary* (2001) pour l'anglais), peu d'ouvrages de référence existent en revanche pour les combinaisons non libres du discours spécialisé autres que les termes complexes nom+nom ou adjectif+nom. La terminologie, qui s'intéresse avant tout à représenter, à travers les termes utilisés pour les désigner, les connaissances propres à un domaine de spécialité, a jusqu'à présent accordé peu d'importance à l'étude du comportement de ces unités dans les textes. Les quelques ouvrages qui documentent les combinaisons lexicales caractéristiques d'un discours spécialisé donnent généralement ces combinaisons à la suite du terme pris comme base d'une combinaison

donnée, sans distinguer clairement les différents types de combinaisons (selon le mode de composition plus ou moins régulier de l'association).

Le manque de ressources terminologiques est d'autant plus critique que les développements rapides des sciences, notamment des sciences et des technologies de l'information, s'accompagnent de développements tout aussi rapides de termes et d'expressions spécialisées : en bio-informatique par exemple, des combinaisons usuelles telles que *genes immigrate*, *[to] explore genomes* ou *[to] edit DNA sequences* sont tout aussi importantes à maîtriser que les termes clés *genome* ou *DNA sequence*[2].

1.3 Repérage automatique des collocations

Pour faciliter le travail des professionnels du discours spécialisé (traducteurs, terminologues, mais aussi personnes travaillant à la localisation[3] de logiciels ou de pages Web), des programmes ont été développés qui identifient automatiquement les collocations d'un domaine de spécialité à partir des textes de ce domaine.

Dans les premiers programmes développés, seuls les outils de l'analyse statistique étaient appliqués au repérage des combinaisons usuelles d'un discours spécialisé. Ces programmes portaient sur des

---

[2] Expressions relevées sur le site d'un centre de ressources en bioinformatique (http://bioinformatics.org/).

[3] Le fait d'adapter un programme à un contexte culturel et linguistique local (Le Jargon Français).

mots-formes apparaissant ensemble dans un espace textuel court et implémentaient des mesures différentes pour déterminer la force de l'association entre deux cooccurrents donnés (Choueka *et al.* 1983; Church et Hanks 1989; Church *et al.* 1991).

Les résultats obtenus sur ces bases étant difficilement exploitables — les combinaisons ramenées décrivaient différents types de relations lexicales —, les chercheurs ont ensuite ajouté des traitements linguistiques aux mesures statistiques utilisées pour repérer les combinaisons significatives d'un texte. Dans une des premières expériences réalisées (Smadja 1993), des critères linguistiques sont utilisés pour évaluer la qualité des combinaisons extraites sur des bases strictement statistiques. Les approches plus récentes appliquent le traitement linguistique en premier et basent l'extraction des collocations sur des corpus enrichis de données linguistiques. On extrait non plus à partir de textes mais à partir d'analyses de ces textes. Dans ces programmes, les mesures statistiques sont ensuite utilisées pour filtrer les combinaisons atypiques (Lin 1998; Kilgarriff et Tugwell 2001; Goldman *et al.* 2001).

Même s'ils basent l'extraction des collocations sur des données linguistiques (les relations syntaxiques qui existent entre deux unités lexicales), les programmes mentionnés ci-dessus ne fournissent généralement pas d'interprétation sémantique des combinaisons retenues. Les combinaisons ramenées manifestent les différentes contraintes qui régissent la mise en discours des unités lexicales (contraintes sémantiques, lexicales et syntaxiques) et doivent toujours être analysées.

1.4 But et méthodologie de la recherche

Le travail présenté dans ces pages visera donc également à formaliser la notion de collocation en langue de spécialité et plus particulièrement celle de collocatif verbal d'un terme. Dans cette perspective, nous appuierons notre méthode de repérage des collocations sur le modèle de description le plus complet de ces phénomènes, celui des fonctions lexicales de la théorie Sens-Texte. Les fonctions lexicales permettent de représenter l'ensemble des phénomènes collocationnels d'une langue selon les deux axes de la description linguistique.

1. Syntaxique : les fonctions lexicales verbales représentent les trois relations grammaticales majeures — sujet, objet direct, objet indirect.

2. Sémantique : 23 fonctions lexicales permettent de représenter les principaux sens des collocatifs verbaux des termes. Elles peuvent également se combiner entre elles pour accroître encore davantage leur pouvoir d'expression.

Les fonctions lexicales définissant notamment des relations syntaxiques entre les unités lexicales, nous proposons d'extraire les expressions usitées d'un domaine de spécialité à partir de textes analysés représentatifs de ce domaine. Le corpus utilisé couvre le domaine de l'informatique. Il sera analysé par le système de TA Logos (Logos Engine® de GlobalWare AG) qui produira les arbres syntaxiques à partir desquels nous proposons d'acquérir les collocations. Le choix de Logos pour le traitement préalable des textes détermine également le choix de la langue

étudiée dans le cadre de cette recherche : Logos traitant seulement l'anglais et l'allemand en tant que langue source, nous rechercherons les combinaisons des termes clés (noms) d'un domaine spécialisé (celui de l'informatique) dans des textes de langue anglaise du domaine.

Les collocations verbales ne décrivent pas seulement les relations qui existent entre le verbe et ses actants. Elles expriment également des relations sémantiques particulières qui correspondent à un petit ensemble de significations prototypiques et manifestent l'effet d'une contrainte entre le verbe et le terme. Ainsi qu'il a été démontré par les premières expériences de repérage et d'acquisition automatique, cette dernière caractéristique des collocations verbales se manifeste par une certaine distribution des unités lexicales ainsi associées dans les textes. Pour isoler les collocations des combinaisons formées librement, nous appliquerons donc un test statistique aux combinaisons extraites sur des bases syntaxiques.

Bien qu'il fasse ressortir un grand nombre de collocations de la liste des combinaisons extraites, le test statistique ne permet pas de les identifier avec une précision suffisante (sans évaluation manuelle). Seul l'encodage sémantique des combinaisons verbe + nom sous la forme de fonctions lexicales peut nous permettre en dernière analyse d'acquérir automatiquement les collocations de la langue de l'informatique. L'encodage sémantique, qui mesurerait également l'aptitude d'une relation syntaxique particulière à exprimer un sens prototypique du domaine modélisé, pourrait alors se substituer au test statistique. Nous examinerons dans la conclusion les défis posés par un tel encodage.

1.5 Structure de la thèse

Le chapitre qui suit présente les fondements théoriques et pratiques de la recherche. Dans la première section (2.1), nous faisons le point sur les modèles de description des collocations en langue générale. Nous présentons les travaux fondateurs du contextualisme britannique, puis nous enchaînons avec l'évaluation des modèles proposés par Hausmann (1979) et Benson (1986). Nous terminons cet état de l'art des modèles de description linguistique des collocations avec la présentation des fonctions lexicales, les outils de description des collocations de la théorie Sens-Texte (Mel'čuk 1996; Mel'čuk 1998) sur lesquels nous avons basé notre méthodologie de repérage des collocations dans les textes.

Nous poursuivons la présentation des phénomènes collocationnels en langue générale par un tour d'horizon des travaux qui rendent compte de ces phénomènes en terminologie et terminographie (section 2.2). Nous nous intéressons, dans un premier temps, aux travaux qui remettent en question l'approche classique à la description terminologique et proposent des modèles de représentation qui intègrent tous la description systématique de la combinatoire des termes (Frawley 1988; Binon *et al.* 2000; Dancette et L'Homme 2002). Nous présentons ensuite les travaux ayant porté spécifiquement sur la représentation des collocations de la langue de spécialité (Heid et Freibott 1991; Béjoint et Thoiron 1992) et décrivons finalement deux des principaux lexiques de cooccurrents spécialisés (Cohen 1986; Meynard 2000).

Nous terminons l'état de l'art (section 2.3) avec la présentation de programmes d'extraction automatique de collocations, basés sur les seules

statistiques (Choueka *et al.* 1983; Church et Hanks 1989; Church *et al.* 1991) et qui combinent approches statistique et symbolique à l'extraction des collocations (Smadja 1993; Lin 1998; Kilgarriff et Tugwell 2001; Goldman *et al.* 2001).

Le troisième chapitre de la thèse présente le corpus utilisé dans le cadre de cette recherche et sur lequel nous avons basé l'extraction automatique des collocations. Il s'agit d'un corpus spécialisé de 600 000 occurrences dont la majorité des textes proviennent de l'Observatoire de linguistique Sens-Texte (OLST) de l'Université de Montréal. La description du corpus fait l'objet de la première section du chapitre (section 3.1). Dans la deuxième section, nous présentons les étapes de préparation du corpus en vue de l'extraction des collocations. Il s'agit du nettoyage et de l'analyse syntaxique des documents par le logiciel de TA Logos (section 3.2).

Le quatrième chapitre représente le cœur de la thèse. Il présente la méthode que nous avons développée pour identifier les collocations de la langue de l'informatique. Nous décrivons Colex, le programme de repérage des collocations, en nous attachant particulièrement à la description des patrons morphosyntaxiques utilisés pour extraire les relations prédicatives des arbres produits par l'analyseur de Logos (section 4.2). Nous terminons la section consacrée à Colex par une évaluation des résultats obtenus à ce stade de l'extraction.

Nous présentons ensuite le traitement statistique appliqué aux combinaisons extraites par Colex afin d'en accroître la pertinence par rapport aux buts de la recherche (section 4.3). Nous évaluons les performances de deux mesures statistiques pour le filtrage des

combinaisons libres. Nous terminons le quatrième chapitre par la présentation des résultats de l'évaluation statistique.

La conclusion récapitule les principales étapes de la méthode développée dans le cadre de cette recherche. Elle discute aussi les limites de notre travail et les perspectives de recherches futures.

2 Fondements théoriques et état de l'art

Ce chapitre présente les fondements théoriques et pratiques de la recherche. Dans la première section (2.1), nous faisons le point sur les modèles de description des collocations en langue générale. Nous poursuivons la présentation des phénomènes collocationnels en langue générale par un tour d'horizon des travaux qui rendent compte de ces phénomènes en terminologie et terminographie (section 2.2). Nous terminons l'état de l'art avec la présentation des programmes développés pour l'extraction automatique de collocations à partir de textes (section 2.3).

2.1 Modèles lexicologiques et lexicographiques des phénomènes collocatifs

Cette section trace un tableau récapitulatif des modèles d'analyse des collocations en lexicologie et lexicographie générale. Nous commençons par une présentation des travaux de J.R. Firth et M.A.K. Halliday, les représentants les plus importants du contextualisme britannique. Le mouvement contextualiste nous intéresse à plus d'un titre puisqu'il a joué un rôle considérable dans le développement de méthodes et techniques en linguistique appliquée, méthodes qui seront également abordées dans le cadre de ce travail. C'est notamment aux travaux de Firth et de Halliday sur les collocations que nous devons le développement de la linguistique de corpus et de méthodes d'analyse lexicale basées sur le calcul statistique. Ces travaux ont conduit à l'élaboration du projet COBUILD, le premier grand corpus construit à des fins lexicographiques, et à la création

d'un dictionnaire qui a bouleversé la pratique lexicographique anglo-saxonne (Sinclair *et al.* 1970; Sinclair 1987).

Après ce rappel de la contribution de l'école de Londres à l'étude des collocations, nous discutons des propositions avancées dans Hausmann (1979) pour la réalisation d'un dictionnaire de collocations. Ces propositions seront notamment reprises dans le *BBI Combinatory Dictionary of English* (*BBI*) de Benson *et al.* (1997). La discussion du modèle de recensement et de présentation des collocations de Hausmann sera donc suivie d'une présentation du *BBI*.

Nous terminerons ce tour d'horizon des modèles descriptifs des collocations en lexicologie et lexicographie générale par une présentation des fonctions lexicales (FL), les outils développés dans le cadre de la lexicologie explicative et combinatoire pour la description des phénomènes de cooccurrence lexicale (Mel'čuk *et al.* 1995; Mel'čuk 1996; Mel'čuk 1998). Les FL servant de base à notre travail, nous leur consacrons la dernière partie de cette section.

2.1.1 Le contextualisme britannique

À de nombreux égards, le contextualisme britannique est à l'origine de la discussion linguistique sur les collocations. Les contextualistes ont non seulement défini le concept, ils ont également préconisé des méthodes d'analyse de ces phénomènes basées sur les textes et le calcul statistique. Le mouvement contextualiste s'est développé en Angleterre sous l'influence de B. Malinowski et de J.R. Firth, tous deux professeurs à l'Université de Londres. (Malinowski y est nommé professeur de la première chaire

d'anthropologie en 1927, Firth professeur de la première chaire de linguistique générale en 1944.)

Anthropologue et ethnologue d'origine polonaise, Malinowski est considéré comme le fondateur du fonctionnalisme, théorie de la culture selon laquelle chaque aspect de la vie culturelle d'une communauté, coutume, objet matériel, idée ou croyance, remplit une fonction particulière, vitale, à l'intérieur de cette communauté. Véhicule privilégié de la culture d'un peuple, la langue joue elle aussi un rôle essentiel dans le système global de la culture ; elle contribue également au bon « fonctionnement » de la communauté. La théorie fonctionnaliste de la langue, ou contextualisme, s'intéresse donc avant tout à la dimension culturelle de la langue et introduit le concept malinowskien de contexte de situation, ou contexte de culture, dans l'interprétation des phénomènes langagiers. Le contexte de situation figurant au cœur de la doctrine contextualiste, il importe de le définir exactement, en se référant tout d'abord à l'article de dictionnaire suivant :

> **CONTEXTE** (...) **1.** Ensemble du texte qui entoure un mot, une phrase, un passage et qui sélectionne son sens, sa valeur (...) **2.** Ensemble des circonstances dans lesquelles s'insère un fait (...) (*Petit Robert* 1984:378)

Le contexte de situation de la théorie fonctionnaliste correspond donc à la deuxième acception de *contexte* dans l'article du *Petit Robert*. Dans l'approche contextualiste, le concept de contexte recouvre les conditions socio-culturelles, historiques et même biographiques de toute énonciation ainsi que les moyens rhétoriques utilisés dans la production

des énoncés. Pour les contextualistes, le concept de contexte tel que défini ci-dessus est très étroitement lié au concept de texte, la manifestation concrète de l'activité langagière. En effet, seul le texte, qu'il s'agisse d'un texte écrit ou transcrit de l'oral, permet d'appréhender le contexte d'une situation d'énonciation et les moyens d'expression formels mis en œuvre dans cette énonciation. Les contextualistes s'attachent donc à décrire la langue utilisée dans une communauté culturelle ou sociale donnée et basent la description linguistique sur l'analyse seule des textes qu'elle produit.

Dès les premiers essais publiés, les contextualistes s'intéressent à la combinatoire des unités lexicales dans les textes. Dans le chapitre qui introduit pour la première fois le terme (cf. Firth 1957), Firth propose d'intégrer la description des combinaisons lexicales aux méthodes et objets d'étude traditionnels de la linguistique :

> "Just as phonetic, phonological, and grammatical forms well established and habitual in any close social group provide a basis for the mutual expectancies of words and sentences at those levels (...) so also the study of the usual collocations of a particular literary form or genre or of a particular author makes possible a clearly defined and precisely stated contribution to what I have termed the spectrum of descriptive linguistics, which handles and states meaning by dispersing it in a range of techniques working at a series of levels." (Firth 1957:195)

Selon Firth, l'étude des collocations d'un énoncé contribue à l'interprétation globale de cet énoncé, au même titre que ses éléments phoniques ou que les conditions socio-culturelles de sa situation

d'énonciation. Firth n'offre pas de définition formelle de la collocation, définie seulement comme l'association habituelle de deux mots ou plus dans un espace de texte court. Pour Firth et les premiers contextualistes, les collocations permettent avant tout d'analyser le style de certains auteurs. L'étude des collocations de Swinburne, poète de l'ère victorienne, est notamment pour Firth l'occasion de nombreux commentaires sur le goût de cet auteur pour les antithèses (cf. les exemples suivants, donnés par Firth) :

(7) *And dreams of bitter sleep and sweet*
(8) *With footless joy and wingless grief*
And twinborn faith and disbelief
(9) *Till life forget and death remember,*
Till thou remember and I forget.

Les collocations jouent également un rôle extrêmement important dans l'interprétation linguistique. Comme le fait remarquer Firth, elles définissent le contexte des mots et permettent la seule appréhension possible de leur sens, dans leur relation de collocation dans ce contexte (*contexte* correspond ici à la première définition du *Petit Robert*) :

> "Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*. This kind of mutuality may be paralleled in most languages and has resulted in similarities of poetic diction in literatures sharing common classical sources." (Firth 1957:196)

On ne trouvera pas dans Firth une caractérisation plus explicite du sens lexical (en contexte ou hors contexte). On ne trouvera pas non plus de description approfondie de méthodes possibles d'interprétation linguistique. Comme nous l'avons déjà remarqué, le fondateur du mouvement contextualiste s'est d'abord intéressé à la description d'œuvres littéraires. C'est son disciple, M.A.K. Halliday, qui s'est attaché à décrire le modèle envisagé par la théorie contextualiste, modèle qui vise avant tout à expliquer le fonctionnement de la langue dans le contexte d'une situation donnée. La grammaire fonctionnelle développée par Halliday définit plusieurs niveaux d'analyse des énoncés (niveau contextuel, grammatical, phonétique, etc.) et fournit une description fonctionnelle des unités linguistiques à chaque niveau envisagé, chaque description devant contribuer à l'interprétation globale de l'énoncé.

Halliday préconisera par ailleurs un niveau d'analyse autonome, lexical ou collocationnel, pour la description des phénomènes de cooccurrence lexicale. Dans l'article consacré à ce niveau lexical d'analyse des énoncés, Halliday attribue à ce niveau des descriptions et méthodes d'analyse différentes de celles offertes par l'analyse grammaticale. La notion de structure de l'analyse grammaticale est ainsi remplacée par celle de cooccurrence linéaire. Aux catégories grammaticales, l'analyse lexicale substitue les ensembles (ouverts) d'unités lexicales ayant la même probabilité d'occurrence dans un environnement donné. Pour Halliday, il ne s'agit nullement de définir deux types de combinaisons, les combinaisons grammaticales et les combinaisons lexicales, mais plutôt deux façons d'analyser une combinaison donnée :

> "Even where the model recognizes two distinct kinds of pattern, these still represent different properties of the total phenomenon of language, not properties of different parts of the phenomenon; all formal items enter into patterns of both kinds. They are grammatical items when described grammatically, as entering (via classes) into closed systems and ordered structures, and lexical items when described lexically, as entering into open sets and linear collocations." (Halliday 1966:155)

Pour une unité lexicale donnée, c'est l'analyse qui permet de mieux rendre compte de ses propriétés de combinatoire qui est sélectionnée. Ainsi, pour décrire la combinatoire des unités lexicales *a* ou *of* dans les textes, le modèle privilégiera l'analyse grammaticale. (Il est possible de définir la combinatoire de ces deux unités en utilisant les catégories de la grammaire.) Une description lexicale de la combinatoire de *a* ou de *of* est beaucoup plus difficile (et assez peu utile), *a* et *of* se combinant librement avec un très grand nombre d'unités lexicales. Pour un adjectif tel que *strong*, il est bien entendu possible de donner une description lexicale de sa combinatoire, celle-ci étant limitée, contrairement à celle de *of*, à l'ensemble des unités lexicales avec lesquelles l'adjectif cooccurre fréquemment. Halliday propose alors d'intégrer le calcul des probabilités à l'analyse des combinaisons lexicales : la combinatoire d'une unité linguistique sera analysable lexicalement lorsque la probabilité d'occurrence de cette unité dans un environnement donné (dans une relation de collocation particulière) s'écartera de façon significative de sa probabilité d'occurrence absolue :

> "In a lexical analysis it is the lexical restriction which is under focus: the extent to which an item is specified by its collocational environment. This therefore takes into account the frequency of the item in a stated environment relative to its total frequency of occurrence. While *a* and *of* are unlikely to occur in any collocationally generalizable environment with a probability significantly different that their overall unconditioned probabilities, there will be environments such that *strong* occurs with a probability greater than chance." (Halliday 1966:156)

L'analyse statistique des combinaisons lexicales devrait donc permettre de définir l'ensemble des unités lexicales ayant les mêmes propriétés de combinatoire (la même probabilité d'occurrence dans un environnement donné). Halliday fournira une description sommaire de la forme que pourrait prendre une telle analyse :

> "If we consider n occurrences of a given (potential) item, calling this item the 'node', and examine its 'collocates' up to m places on either side, giving a 'span' of $2m$, the $2mn$ occurrences of collocates will show a certain frequency of distribution. For example, if for 2,000 occurrences of *sun* we list the three preceding and three following lexical items, the 12,000 occurrences of its collocates might show a distribution beginning with *bright, hot, shine, light, lie, come out* and ending with a large number of items each occurring only once." (Halliday 1966:158)

Les travaux qui suivront ces premiers essais d'analyse des phénomènes de cooccurrence lexicale ne manqueront pas de souligner les limites de l'approche fonctionnelle des contextualistes. La critique la plus

sérieuse des résultats de l'école contextualiste sera formulée notamment par Hausmann (voir la présentation de la théorie de Hausmann dans la sous-section suivante) : l'analyse statistique des combinaisons lexicales telle que présentée ci-dessus retient toutes les combinaisons observables dans les textes sans qu'il soit possible, en l'absence d'une caractérisation syntaxique ou sémantique des combinaisons retenues, de les différencier davantage[1]. Malgré ses limites, la théorie contextualiste a exercé et continue d'exercer une influence considérable dans l'étude des phénomènes collocationnels. Les travaux des contextualistes ont particulièrement influencé le développement de la linguistique de corpus et de méthodes d'analyse basées sur les statistiques. Ces travaux sont ainsi à la base des dictionnaires COBUILD, les premiers dictionnaires entièrement réalisés à partir de corpus (Sinclair *et al.* 1970). Rappelons pour finir la contribution des contextualistes à l'étude des registres de la langue (l'importance accordée à l'analyse des combinaisons lexicales), contribution à laquelle ne peut manquer d'être redevable toute recherche en langues de spécialité et, par conséquent, la présente recherche.

2.1.2 La théorie de Hausmann

Les contextualistes britanniques ont engagé la discussion linguistique moderne sur les collocations. Parmi les recherches que

---

[1] À la fin de son article sur le niveau lexical d'analyse du texte, Halliday souligne cependant l'opportunité d'une analyse lexicogrammaticale des collocations, c'est-à-dire d'une analyse intégrant l'examen de la structure grammaticale des collocations.

suscitèrent les travaux du contextualisme, celles du linguiste et lexicographe allemand F. J. Hausmann demeurent fondamentales. Spécialiste du français, Hausmann s'intéresse depuis longtemps à la description des expressions de cette langue. Dans l'exposé qu'il présente en 1979 à la Faculté des Lettres modernes de l'Université de Strasbourg (qui sera publié sous le titre *Un dictionnaire des collocations est-il possible ?*), Hausmann réagit aux résultats des contextualistes, résultats sur lesquels il est impossible, selon lui, de fonder la description des collocations d'une langue. Les combinaisons lexicales que définit l'analyse statistique de Halliday sont en effet bien trop nombreuses pour que l'on puisse envisager de les rassembler toutes dans un dictionnaire. Ces combinaisons comptent également nombre d'expressions parfaitement compréhensibles ne méritant pas de faire l'objet d'une inscription particulière. La construction d'un dictionnaire des collocations du français comblerait cependant les attentes de tous ceux qui cherchent à bien s'exprimer dans cette langue. L'exposé de Strasbourg sera donc pour Hausmann l'occasion de jeter les bases théoriques et pratiques d'un tel dictionnaire.

L'intérêt du public pour un ouvrage qui rassemble les expressions courantes du français ne date pas d'hier. Ainsi que le rappelle Hausmann, il existe en France une longue tradition de compilation de dictionnaires de collocations, tradition qui commence en 1571 avec la parution des *Epithètes de M. de La Porte, Parisien.* L'ouvrage, qui fera l'objet de six rééditions (la dernière édition date de 1612), contient 80 000 expressions du français du XVI^{e} siècle, réparties dans 4 000 articles substantifs.

L'auteur y recense les combinaisons noms+adjectifs de la littérature de son temps. D'après Hausmann, peu d'entre elles sont parvenues jusqu'à nous.

Plusieurs auteurs reprendront l'entreprise de Maurice de La Porte aux XVII^e^ et XVIII^e^ siècles et produiront à leur tour des ouvrages destinés à ceux qui sont à la recherche du mot juste, ou, selon l'un des successeurs de La Porte, « aux poètes, aux orateurs, aux jeunes gens qui entrent dans la carrière des sciences, et à tous ceux qui veulent écrire correctement, tant en vers qu'en prose ». Telle est également la vocation du dernier dictionnaire de collocations mentionné par Hausmann, dictionnaire adressé aux élèves des écoles primaires et publié sous le titre *Les mots et les idées. Dictionnaire des termes cadrant avec les idées* (Lacroix 1958). Cet ouvrage, le plus proche de nous puisque la dernière réimpression, aujourd'hui épuisée, date de 1967, est également le dernier dictionnaire consacré entièrement aux collocations du français : des ouvrages plus récents tels que les *Dictionnaire des expressions et locutions* (Rey & Chantreau 1999) et *Dictionnaire des expressions et locutions traditionnelles* (Rat 1999) s'attachent davantage à la description des expressions idiomatiques du français et ne contiennent qu'une petite partie des collocations que l'on trouve dans le dictionnaire d'Ulysse Lacroix[2].

---

[2] Nous devons également mentionner le dictionnaire analogique de Paul Rouaix (1989) destiné lui aussi à l'enseignement du vocabulaire ainsi que le *Dictionnaire des cooccurrences* de Jacques Beauchesne (2001), traducteur et terminologue au gouvernement fédéral du Canada. À bien des égards, ce dernier dictionnaire, produit de trente années de lecture, constitue une version considérablement enrichie de l'ouvrage conçu par Ulysse Lacroix en 1958.

Avec plusieurs milliers d'articles vedettes (composés presque entièrement d'unités nominales), *Les mots et les idées* est une ressource extrêmement riche qui indique, pour chaque nom, non seulement les adjectifs, mais aussi les verbes avec lesquels il se combine couramment. Les collocatifs verbaux d'un article vedette sont organisés par ailleurs selon la fonction grammaticale (sujet ou objet) du nom à l'intérieur de la combinaison. Nous reproduisons ci-dessous l'article de DOUTE dans *Les mots et les idées* (1958).

> Doute. Émettre, formuler, conserver, éclaircir, dissiper, élever un doute. Un doute s'élève, naît, surgit, subsiste, persiste, plane. — QUAL. : affreux, léger, subit.

Figure 2-1 Article de DOUTE (*Les mots et les idées*)

Le relevé systématique et la présentation des collocations des unités nominales du français dans le dictionnaire de Lacroix contraste de façon marquée avec la façon de présenter le même matériel dans d'autres dictionnaires du français, notamment le *Petit Robert*. Nous reproduisons ci-dessous l'article de DOUTE du *Petit Robert*. (Nous avons abrégé cet article pour les besoins de l'exposition et signalé les collocations par des italiques.)

DOUTE ♦ **1°** État de l'esprit qui doute (...) *Être dans le doute au sujet de qqch. Laisser qqn dans le doute. — Mettre une assertion en doute.* ♦ **2°** UN DOUTE : jugement par lequel on doute de qqch. *Avoir un doute sur l'authenticité d'un document, sur la réussite d'une affaire. Laisser planer un doute. Lever, éclaircir, dissiper un doute.*

Figure 2-2 Article de DOUTE (*Petit Robert*)

Comme il ressort de l'exemple précédent, l'information collocationnelle donnée par les dictionnaires de langue tels que le *Petit Robert* est rare et difficile à trouver : les collocations sont souvent présentées dans les exemples. Hausmann souligne également qu'un grand nombre de collocations sont indiquées sous les collocatifs. Toujours dans le *Petit Robert*, les collocations *parfum léger, taille légère, bruit léger* ou *blessure légère* figurent toutes sous l'article de LÉGER et non sous le nom correspondant. Pour Hausmann, l'utilité d'un dictionnaire qui présenterait logiquement les collocations du français ne fait aucun doute. Il importe auparavant de proposer une définition de la collocation qui tienne compte des caractéristiques formelles de cette unité linguistique et permette également d'éliminer la majorité des combinaisons lexicales relevées dans les textes (que retiendra l'analyse lexicale des contextualistes).

Cette dernière considération permet d'isoler ce qui constitue aux yeux de Hausmann la caractéristique principale de la collocation : plus que la combinaison probable de deux unités lexicales, la collocation est une combinaison contrainte. Les unités lexicales sont associées de façon conventionnelle sans que ne puisse intervenir la créativité des locuteurs. Pour Hausmann, c'est la nature essentiellement arbitraire du lien

collocationnel qui permet de distinguer la collocation des autres combinaisons de la langue.

Hausmann identifie ensuite une deuxième caractéristique, tout aussi importante : la collocation n'est pas seulement une combinaison sous contrainte, c'est également une combinaison orientée. C'est cette orientation qui permet d'appeler l'un des membres la base de la collocation et l'autre le collocatif :

> "En effet, dans la collocation *célibataire endurci*, le signifié de la base (*célibataire*) est autonome. La base n'a pas besoin du collocatif (*endurci*) pour être clairement définie. Il en va tout autrement pour le collocatif qui ne réalise pleinement son signifié qu'en combinaison avec une base (*célibataire, pécheur, âme*, etc.)." (Hausmann 1979:191)

La différence de statut entre base et collocatif apparaît clairement dans un dictionnaire de langue tel que le *Petit Robert*. Comme nous l'avons déjà remarqué avec l'exemple de LÉGER, les dictionnaires traditionnels ne peuvent définir entièrement le sens d'un collocatif sans recourir à la dimension syntagmatique des collocations, alors qu'ils peuvent s'en passer pour les bases. C'est pour cette raison que les collocations sont souvent données dans les articles consacrées aux collocatifs, et non dans ceux des bases.

La différence entre base et collocatif est essentielle en apprentissage dans la mesure où les bases peuvent s'apprendre isolément alors que les collocatifs ne peuvent s'apprendre que dans l'entourage de leurs bases. Finalement, la différence entre les deux membres d'une collocation est essentielle dans une perspective de production d'énoncés, la perspective

donnée par Hausmann au dictionnaire des collocations du français : celui qui rédige un texte cherche le collocatif à partir d'une base connue et non inversement. Il reste donc à définir les unités de base du français qui formeront la nomenclature du futur dictionnaire. Pour Hausmann, il ne peut s'agir que d'unités nominales, « la partie du discours le plus près du monde des choses et des êtres », autour de laquelle le locuteur est censé organiser sa pensée. Le dictionnaire de langue écrite préconisé par Hausmann, pour répondre aux besoins de ses utilisateurs éventuels, doit donc organiser le matériel collocationnel (seulement ce qui montre une contrainte arbitraire) autour des bases nominales du français.

Dans son organisation, le dictionnaire de Hausmann suit le modèle de Lacroix (1958) (macrostructure essentiellement constituée de noms). Hausmann ajoute deux points importants. Il suggère de combler les lacunes au niveau de la microstructure, les articles de Lacroix ne donnant que les collocatifs adjectivaux et verbaux des noms étudiés. Il propose également d'inclure une caractérisation syntaxique et sémantique des collocations présentées dans le dictionnaire. Plus spécifiquement, Hausmann établit le cadre des collocations possibles d'une base nominale française, cadre dans lequel il fait entrer les constructions base + verbe, verbe + base, verbe + préposition + base, (adjectif) + base + (adjectif) et nom + préposition + base. Finalement, Hausmann propose de développer un système sémantique simple afin de catégoriser les concepts énoncés à partir de chaque base. Pour l'article de DOUTE présenté un peu plus tôt, Hausmann propose l'ébauche suivante :

1. [N+v] **NAITRE**, **EXISTER** : naître, surgir, m'envahit, plane, subsiste, persiste ; **DISPARAITRE** : s'évanouir, s'envoler. 2. [v+N] **AVOIR** : avoir, concevoir, éprouver, il me vient des doutes ; **FAIRE NAITRE** : inspirer ; **EXPRIMER** : émettre, formuler ; **FAIRE DISPARAITRE** : lever, écarter, éclaircir, dissiper, balayer. 3. [v+prép+N] (être) assailli de doutes, rongé, tourmenté par le doute ; être, laisser dans le doute ; mettre, révoquer en doute. 4. [(a)+N+(a)] : légers -, - affreux, subits, persistants, bien fondés. 5. [N+prép+N] : le supplice du doute.

Figure 2-3 Article de DOUTE élaboré par Hausmann (1979)

On voit s'ébaucher ici une méthode de description des collocations qui tient compte à la fois de la structure syntaxique et de la sémantique propres à ces phénomènes : les mots en caractères gras désignent les éléments du système sémantique développé pour catégoriser les sens collocationnels. Les propositions concrètes de Hausmann trouveront un écho dans le *BBI Combinatory Dictionary of English* (*BBI*) de Benson *et al.* (1997). Nous allons donc examiner maintenant le *BBI*.

2.1.3 La description lexicographique du *BBI*

Le *BBI* est le dictionnaire de collocations de l'anglais le plus complet qui existe. Bien qu'il ait été conçu dans l'ignorance complète des travaux de Hausmann (Benson 1989), le *BBI* représente, par son contenu et sa forme, une mise en application remarquable des principes présentés par celui-ci dans l'exposé de Strasbourg (1979). Les objectifs poursuivis dans les deux cas peuvent expliquer les liens étroits qui existent entre les travaux de ces différents chercheurs. Le *BBI* se présente en effet comme un dictionnaire d'apprentissage de l'anglais qui, tel le dictionnaire de langue préconisé par Hausmann, doit avant tout aider ses utilisateurs à

s'exprimer le plus naturellement possible dans cette langue. Il tente donc de répondre aux questions soulevées par Hausmann quelques années plus tôt : quelles expressions un dictionnaire d'apprentissage doit-il répertorier et comment doit-il les présenter ?

Le *BBI* s'inscrit dans une longue tradition de dictionnaires d'apprentissage de l'anglais. Le premier de ces dictionnaires, le *Oxford Advanced Learner's Dictionary of Current English* (*OALDCE*), paraît pour la première fois en Angleterre en 1948. Il est l'œuvre du professeur d'anglais et lexicographe A. S. Hornby, qui reprend les travaux commencés par Harold Palmer — le premier à s'être intéressé à la description des expressions usuelles de l'anglais dans un ouvrage destiné aux apprenants de cette langue. Palmer développera notamment un système de codification de la valence syntaxique des verbes anglais. Le *OALDCE* devient rapidement un ouvrage indispensable pour l'enseignement de l'anglais et connaît de nombreuses rééditions, la dernière datant de 2000 (Hornby 2000). Un autre dictionnaire d'apprentissage apparaît plus récemment et s'impose comme un rival sérieux du *OALDCE*. Il s'agit du *Longman Dictionary of Contemporary English* (*LDOCE*). Ce dictionnaire se propose de combiner le traitement de l'information grammaticale et collocationnelle des entrées, caractéristiques de la tradition Palmer-Hornby, avec l'utilisation d'un vocabulaire définitoire. Le *LDOCE* développe encore plus l'information fournie dans le *OALDCE*, en ajoutant notamment la description des propriétés syntaxiques des noms, adjectives et adverbes. Il paraît pour la première fois en 1978 (la dernière édition de poche date de 1996).

Pour les auteurs du *BBI*, c'est le traitement presque exclusif de l'information grammaticale qui constitue le problème le plus sérieux des deux dictionnaires dont il vient d'être question. Les dictionnaires d'apprentissage tels que le *OALDCE* et le *LDOCE* privilégient en effet la description de la combinatoire grammaticale des entrées qu'ils contiennent (nombre et type de compléments régis) et donnent finalement assez peu de collocations. Nous reproduisons à titre d'exemple l'article de DOUBT du *LDOCE*.

doubt[1]
S1, W1
noun

1 ▶UNCERTAIN FEELING◀ [countable, uncountable] a feeling or feelings of being uncertain about something
[+ about/as to]: *Maisie expressed private doubts about Lawrence's sanity.*
[+ whether/who/what etc]: *There's no doubt who was responsible for this outrage.*
[+ (that)]: *I have little doubt that the coup will succeed.* | cast doubt(s) on sth/raise doubts about sth (=say that something may not be true or real): *The new evidence cast some doubt on his reliability as a witness.* | an element of doubt (=a slight doubt) | without a shadow of a doubt (=there is no doubt at all)

Figure 2-4 Article de DOUBT (*LDOCE*)

On voit dans cet exemple que le *LDOCE* consacre une grande partie de l'article d'une entrée à la description de sa combinatoire grammaticale (représentée entre crochets). Les collocations sont en revanche peu nombreuses (*[to] cast doubt(s) on sth, [to] raise doubts about sth*) et présentées à l'intérieur d'exemples (*Maisie expressed private doubts about Lawrence's sanity*) ou de précisions définitoires (*a slight doubt*).

La présentation des collocations (à l'intérieur des exemples et du matériel définitoire) constitue donc le deuxième problème des dictionnaires

d'apprentissage existants pour les auteurs du *BBI*. Les collocations sont difficilement repérables, d'autant plus qu'elles sont généralement indiquées dans les articles consacrés aux verbes, l'unité lexicale problématique pour l'apprenant qui cherche à produire une expression à partir d'une base nominale connue. Par exemple, le *LDOCE* donne les collocations *[to] draw attention*, *[to] draw a crowd*, *[to] draw a gun* à l'entrée du verbe mais ne les indique pas dans les articles des noms correspondants (Benson 1989).

Finalement, Benson et ses collègues reprochent aux dictionnaires de collocations existants, et particulièrement à ceux rédigés à l'étranger, de donner un très grand nombre de combinaisons libres, parfaitement prévisibles selon le sens de leurs composantes individuelles, et ne posant donc pas de difficulté particulière pour les apprenants. Toutes ces constatations vont motiver la construction d'un nouveau dictionnaire de collocations qui servira également à clarifier une notion encore assez mal comprise à l'époque de la parution du *BBI*.

Comme pour Hausmann, la construction du *BBI* va donc s'accompagner de la recherche et de l'élaboration d'une définition des collocations. Une première définition est proposée dans Benson (1985). Elle offre une caractérisation sémantique des collocations qui les différencie à la fois des expressions idiomatiques (ou expressions figées) et des combinaisons libres. Les collocations que Benson définit alors comme des combinaisons faiblement figées (*loosely fixed combinations*) se distinguent des combinaisons libres de deux façons :

1. par une synonymie limitée — on peut rarement substituer un synonyme au verbe employé dans une collocation donnée — et

2. par un usage fréquent — les collocations viennent spontanément à l'esprit des locuteurs qui les emploient de façon régulière.

Dans un article publié quelques années plus tard (1989), Benson insistera davantage sur le caractère arbitraire, non prévisible de ces associations, caractère qui s'appréhende le plus directement dans l'étude contrastive de deux langues. C'est cette dernière caractéristique qui rend la collocation particulièrement difficile à maîtriser pour un étranger. Elle joue donc un rôle primordial dans la sélection des combinaisons retenues dans le *BBI*.

Les auteurs du *BBI* vont ainsi s'attacher à décrire les combinaisons verbe + nom formées à partir de verbes qui dénotent la création ou l'animation (*Creation or Activation* = CA) ou la disparition ou l'annulation (*Eradication or Nullification* = EN). Selon Benson, ces combinaisons sont majoritairement arbitraires en anglais et ne peuvent donc être produites spontanément par les étrangers. Nous donnons ci-dessous des exemples de collocations CA et EN données dans le *BBI* :

- collocations CA : *[to] compile a dictionary, [to] compose music, [to] set an alarm, [to] launch a missile, [to] perform an operation, [to] issue a warning*, etc.

- collocations EN : *[to] lift a blockade, [to] dispel fear, [to] demolish a house, [to] revoke a license, [to] withdraw an offer, [to] crush resistance*, etc.

Le *BBI* n'inclura cependant pas les combinaisons CA ou EN qui sont composées librement. Des exemples de ces combinaisons libres sont données par les auteurs du *BBI.* Il s'agit selon eux de combinaisons formées de verbes pouvant se combiner avec un très grand nombre de noms sémantiquement apparentés : *[to] build bridges (houses, roads), [to] cause damage (deafness, a death), [to] cook meat (potatoes, vegetables), [to] grow apples (bananas, corn)*, etc. Pour Benson, le mode de composition régulier de ces expressions les rend parfaitement accessibles aux étudiants de l'anglais qui les produisent généralement sans l'aide d'aucun dictionnaire.

Le caractère régulier ou irrégulier d'une combinaison donnée, mesuré en fonction de la difficulté pour un apprenant à la produire spontanément, détermine finalement la sélection de toutes les combinaisons lexicales décrites dans le *BBI.* Il permet ainsi de retenir les combinaisons verbe + nom non prévisibles qui n'entrent pas dans la catégorie des collocations CA ou EN (*[to] do the laundry, [to] decline a noun, [to] take one's seat*, etc.) et justifie également l'inclusion des combinaisons lexicales qui décrivent d'autres types de relations syntaxiques à l'intérieur

de ce dictionnaire. Nous donnons ci-dessous des exemples pour chacun des types de combinaisons retenus[3] :

- adjectif + nom : *reckless abandon*, *crushing defeat*, *rough estimate*, *sweeping generalization*, etc.
- nom + verbe : *adjectives modify*, *alarms go off*, *bees buzz*, etc.
- nom + nom : *a colony of bees*, *a bit of advice*, etc.
- adverbe + adjectif : *deeply absorbed*, *sound asleep*, etc.
- adverbe + verbe : *affect deeply*, *amuse thoroughly*, etc.

La présentation des combinaisons lexicales de l'anglais dans le *BBI* respecte elle aussi les principes de Hausmann concernant le placement des collocations à l'intérieur des dictionnaires d'apprentissage. Les combinaisons ayant une base nominale (la majorité des collocations recensées dans le dictionnaire) sont indiquées systématiquement à l'entrée du nom. (Les collocations adverbe + adjectif et adverbe + verbe sont stockées sous l'adjectif et le verbe respectivement.) L'ordre des présentation des collocations à l'intérieur des articles est le suivant :

---

[3] La typologie syntaxique et sémantique des collocations retenues dans le *BBI* est basée dans une large mesure sur les fonctions lexicales développées par Igor Mel'čuk et ses collègues lors de la création du dictionnaire explicatif et combinatoire du russe (Apresyan *et al.* 1969) (cf. Benson 1985; Benson *et al.* 1986).

verbe + nom (CA), verbe + nom (EN), adjectif + nom, nom + verbe, nom + nom. Nous reproduisons ci-dessous l'article de DOUBT du *BBI*.

doubt I *n.* 1. to plant; raise (a) ~ (her proposal raised serious ~s in my mind) 2. to cast ~ on 3. to feel ~; to entertain, harbor, have ~s about 4. to express, voice (a) ~ 5. to clear up, dispel, resolve a ~ 6. (a) deep, serious, strong; gnawing; lingering; reasonable; slight ~ 7. ~s appear, arise 8. a ~ about, of 9. (a) ~ that + clause (he expressed serious ~ that he could finish the job on time) 10. beyond (a shadow of) a ~; without a ~ 11. in ~ (the result was never in serious ~) 12. (misc.) to give smb. the benefit of the ~; (colloq.) there is no ~ about it: she's the best

Figure 2-5 Article de DOUBT (*BBI*)

Ainsi qu'il a été signalé au début de cette section, le *BBI* est un dictionnaire unique en son genre, qui offre la liste la plus complète de combinaisons lexicales de l'anglais qui existe. Dans la dernière version publiée, 18 000 entrées et 90 000 collocations sont recensées. Le *BBI* suit la pratique des dictionnaires d'apprentissage de l'anglais et donne également la combinatoire grammaticale des entrées. (Elle est indiquée après la combinatoire lexicale.)

Le travail accompli par les auteurs de ce dictionnaire est remarquable également du point de vue de la contribution théorique et méthodologique à l'étude des collocations. Le recensement des collocations de l'anglais s'est accompagné d'un véritable effort de formalisation de ces phénomènes et de l'élaboration d'une définition opérationnelle qui permet de distinguer les collocations des autres expressions de la langue,

particulièrement des combinaisons libres. Le *BBI* fut également l'un des premiers dictionnaires commerciaux à offrir une classification formelle des collocations recensées selon les différents types de relations syntaxiques qu'elles mettent en jeu. On peut seulement regretter que le dictionnaire n'ait pas inclus une caractérisation des propriétés sémantiques des collocations, caractérisation jugée également nécessaire par Hausmann. Dans la version publiée du *BBI*, la description des collocations verbe + nom ne reprend pas les symboles CA ou EN qui constituaient pourtant un début de classification sémantique. Une caractérisation sémantique existe cependant, de façon implicite, dans le regroupement des sens synonymiques des collocatifs à l'intérieur des articles. Ces regroupements sont signalés par des numéros dans le cas des combinaisons verbe + nom et par des points virgules dans le cas des autres combinaisons (cf. l'article de DOUBT ci-dessus).

Nous abordons maintenant le dernier modèle proposé pour la représentation des collocations dont il sera question dans la présente recherche. Il s'agit également du modèle le plus complet.

2.1.4 Les fonctions lexicales de la théorie Sens-Texte

2.1.4.1 Présentation de la notion de fonction lexicale

Le dernier modèle présenté ici, et sur lequel nous avons établi les patrons de repérage des collocations verbe + nom de l'informatique, représente une contribution fondamentale à la lexicographie en général et à l'étude des collocations en particulier. Il s'agit des fonctions lexicales, élaborées dans le cadre de la théorie linguistique Sens-Texte. Le système

des fonctions lexicales a été mis au point dans les années 60 par un groupe de chercheurs réunis autour du linguiste Igor Mel'čuk alors qu'ils travaillaient à la construction d'un dictionnaire explicatif et combinatoire (DEC) du russe (Apresyan *et al.* 1969). Le DEC du russe devait notamment permettre la génération automatique de textes dans cette langue. Les travaux commencés par Mel'čuk sur ce nouveau type de dictionnaire se poursuivent aujourd'hui à l'Université de Montréal avec la construction d'un DEC du français (4 volumes publiés), l'élaboration d'une version électronique du DEC (le DiCo) et d'une version grand public, utilisant un paraphrasage linguistique des liens de fonctions lexicales, le *Lexique Actif du Français* (LAF) (cf. Polguère 2000).

Le DEC est au cœur du modèle linguistique Sens-Texte. La description qu'il offre du lexique d'une langue soutient l'opération principale de cette langue selon la théorie Sens-Texte, la traduction d'une représentation sémantique donnée dans les multiples textes ou énoncés synonymiques qui en sont l'expression linguistique. Le DEC offre non seulement la description des significations lexicales d'une langue, il décrit également toutes les relations qui existent entre les unités lexicales de cette langue et qui sont mises en jeu dans la production des énoncés. Les relations lexicales qui soutiennent la traduction Sens-Texte se manifestent selon deux axes : l'axe paradigmatique, qui réunit les unités lexicales entretenant un lien sémantique fort et l'axe syntagmatique, qui décrit la combinatoire des unités lexicales à l'intérieur des phrases.

Nous avons rencontré des exemples de relations de combinatoire lorsque nous avons étudié les collocations de la lexie DOUBT dans l'article du *BBI* qui lui est consacré *([to] feel, [to] entertain, [to] harbor, [to] have*

doubts, etc.). DOUBT est également reliée aux lexies de l'anglais qui possèdent plus ou moins le même sens qu'elle (UNCERTAINTY, HESITATION, APPREHENSION) et à celles qui possèdent des sens opposés (CERTAINTY, CONFIDENCE, TRUST). Une partie du réseau lexical (très vaste) de DOUBT est représentée dans la phrase suivante :

(10) *The articles of impeachment said that Nixon "has raised substantial doubt as to his judicial integrity, undermined confidence in the integrity and impartiality of the judiciary, betrayed the trust of the people of the United States and brought disrepute on the federal courts and the administration of justice."*

Les relations lexicales dont il vient d'être question sont attestées dans toutes les langues naturelles modélisées dans le cadre de la théorie Sens-Texte. Elles sont représentées dans la TST par le même outil formel, appelé fonction lexicale[4]. Il s'agit d'une fonction au sens mathématique qui admet une unité lexicale comme argument et qui rend un ensemble d'unités lexicales comme valeur. Elle est représentée par la formule traditionnelle $f(x) = y$.

---

[4] La liste des fonctions lexicales standard a été établie à la suite de l'examen de tous les patrons récurrents trouvés dans un nombre considérable de langues appartenant à des familles typologiques fort distinctes.

Les fonctions lexicales sont formellement définies de la façon suivante[5].

> Une fonction lexicale **f** décrit une relation existant entre une lexie L — l'argument de **f** — et un ensemble de lexies appelé la valeur de l'application de **f** à la lexie L. La fonction lexicale **f** est telle que :
> 1. l'expression **f**(L) représente l'application de **f** à la lexie L ;
> 2. chaque élément de la valeur de **f**(L) est lié à L de la même façon.

Nous donnons tout de suite des exemples de fonctions lexicales pour illustrer cette définition :

- **Syn** servira à décrire la relation entre DOUBT et UNCERTAINTY, HESITATION, APPREHENSION ;
- **Anti** permettra de modéliser celle qui existe entre DOUBT et CERTAINTY, CONFIDENCE, TRUST ;
- **Magn** décrira la relation syntagmatique entre DOUBT et DEEP, SERIOUS, STRONG, etc. et
- **Oper** celle entre DOUBT et FEEL, ENTERTAIN, HARBOR.

---

[5] La définition formelle des fonctions lexicales ainsi que la présentation détaillée qui en est faite un peu plus loin dans ces pages est basée sur Polguère (2003a). Il s'agit de la présentation la plus accessible des fonctions lexicales et qui met également en relief la nature essentiellement pratique de ces dernières.

Les fonctions lexicales formalisent des liens qui existent entre un très grand nombre d'unités lexicales. Elles ont elles-mêmes des significations générales, qui appartiennent à un petit ensemble de significations de base, très proches de primitifs sémantiques : 'non' pour **Anti**, 'grand' pour **Magn**, 'faire' pour **Oper**. Dans le cas des fonctions lexicales syntagmatiques, ces significations sont liées à des rôles syntaxiques particuliers : **Magn** décrit un adjectif ou un adverbe modificateur, **Oper**, un verbe sémantiquement vide qui « verbalise » le nom qui remplit auprès de lui le rôle de complément d'objet[6].

Les fonctions lexicales ne sont pas pour autant des unités lexicales de la langue mais correspondent plutôt à des métalexies, qui ne peuvent être appréhendées qu'en pratique, lorsqu'elles sont appliquées à une unité lexicale donnée. La description des fonctions lexicales qui vient d'être faite couvre le type le plus important de fonctions lexicales, les fonctions lexicales standard simples. C'est également ce type de fonctions qui est décrit plus en détail dans le reste de cette section. La description des fonctions lexicales standard simples sera illustrée à l'aide de l'unité lexicale du français DOUTE.

2.1.4.2 FL paradigmatiques

Les fonctions lexicales paradigmatiques modélisent les relations qui existent entre lexies connectées à l'intérieur d'un même paradigme

---

[6] Une présentation du caractère général des fonctions lexicales est donnée dans Polguère (2003b).

sémantique et qui peuvent, dans certaines circonstances, se substituer l'une à l'autre pour dénoter une situation donnée. Plus spécifiquement, elles modélisent les trois types de relations lexicales suivantes :

1. relations sémantiques fondamentales (synonymie, antonymie, hyperonymie),
2. dérivés syntaxiques (nominalisations, etc.) et
3. dérivés sémantiques.

Nous présentons des exemples de ces trois types de fonctions lexicales dans les pages qui suivent.

2.1.4.2.1 Synonyme

La première fonction lexicale, **Syn**, modélise la relation sémantique première, centrale pour la théorie Sens-Texte puisque c'est sur elle que repose le système de paraphrasage.

Syn associe une unité lexicale L à l'ensemble de ses synonymes. Ces synonymes sont le plus souvent des synonymes approximatifs de L, les synonymes exacts étant rarissimes dans la langue. La fonction lexicale **Syn** permet donc de représenter trois types de synonymes approximatifs :

Syn(*claque*) = **fam.** *baffle*
Syn$_{\supset}$(*doute*) = *incrédulité*

Le sens de INCRÉDULITÉ est plus riche que celui de DOUTE (il inclut celui de DOUTE).

Syn$_{\subset}$(*incrédulité*) = *doute*

Le sens de DOUTE est moins riche que celui de INCRÉDULITÉ (il est inclus dans celui de INCRÉDULITÉ).

Syn$_{\cap}$(*doute*) = *perplexité*

Les sens de DOUTE et de PERPLEXITÉ ont une intersection très importante, mais se distinguent tout de même par certaines composantes.

2.1.4.2.2 Antonyme

Cette fonction lexicale lie des unités lexicales qui sont elles aussi sémantiquement très proches, les antonymes se différenciant uniquement par la négation d'une de leurs composantes sémantiques. Ainsi DOUTE et CERTITUDE sont des antonymes. Ces deux unités lexicales s'opposent sémantiquement comme le démontre l'examen de leurs signifiés[7] :

CERTITUDE ≅ 'état d'esprit de X à propos de Y selon lequel X, qui n'a pas de preuves réelles de l'existence de Y, pense que Y existe'

DOUTE ≅ 'état d'esprit de X à propos de Y selon lequel X, qui n'a pas de preuves réelles de la non-existence de Y, pense que Y n'existe pas'

---

[7] La définition proposée ici pour DOUTE (et pour CERTITUDE) identifie les actants sémantiques X et Y de ce prédicat.

Comme dans le cas des synonymes, la théorie Sens-Texte distingue des antonymes exacts et approximatifs. Nous indiquons ci-dessous d'autres quasi-antonymes de DOUTE.

Anti$_{\supset}$(*doute*) = *confiance, conviction*

Anti$_{\subset}$(*doute*) = *assurance*

2.1.4.2.3 Dérivés syntaxiques

Le deuxième groupe de fonctions lexicales paradigmatiques modélise les liens de dérivation syntaxique. Ces fonctions relient entre elles des lexies synonymiques qui appartiennent à des parties du discours différentes.

S$_0$ associe un équivalent nominal (un substantif) à un verbe, adjectif ou adverbe.

S$_0$(*douter*) = *doute*

V$_0$ associe à un nom, adjectif ou adverbe, un équivalent verbal.

V$_0$(*doute*) = *douter*

2.1.4.2.4 Dérivés sémantiques

Le troisième groupe de fonctions lexicales paradigmatiques associe à une unité lexicale prédicative telle que DOUTE l'ensemble de ses dérivés

sémantiques[8]. Ici encore, les unités lexicales partagent une composante importante de sens. Contrairement aux dérivés syntaxiques vus précédemment, les dérivés sémantiques d'une unité lexicale ajoutent quelque chose au sens de celle-ci : ils incluent le sens de l'unité lexicale dont ils sont dérivés.

S$_i$ relie une unité lexicale prédicative L au nom standard de son $i^{ème}$ actant sémantique. L'indice qui apparaît après le nom de la fonction lexicale identifie cet actant.

S$_1$(*conduire*) = *conducteur*

A$_i$ relie une unité lexicale prédicative L au modificateur standard de son $i^{ème}$ actant ; le modificateur exprime auprès de l'actant le sens 'tel que celui-ci est dans la situation dénotée par L'.

A$_1$(*doute*) = *pris* [*de* ~], *dans* [*le* ~]
A$_2$(*doute*) = *dans* [*le* ~] // *douteux*

Les deux derniers groupes de fonctions lexicales (dérivés sémantiques et syntaxiques) relient entre elles des lexies d'une façon qui rappelle la dérivation morphologique. Les liens sont toutefois établis sur des bases sémantiques (même dans le cas des dérivés syntaxiques). Ils ne s'accompagnent donc pas nécessairement d'une dérivation morphologique

---

[8] Rappelons qu'un prédicat sémantique est une unité lexicale qui dénote une situation impliquant au moins un participant qui est un actant sémantique, DOUTE, AMOUR, SOMMEIL parmi les unités nominales mais aussi ASSIETTE (utilisée par qqn. pour qqch.) ou NEZ (de qqn.).

correspondante (cf. les exemples de dérivation $\mathbf{A_1}$(*doute*) donnés ci-dessus). C'est pour cette raison que les liens lexicaux modélisés par ces deux derniers groupes de fonctions lexicales sont décrits dans la théorie Sens-Texte comme des dérivations sémantiques.

2.1.4.3 FL syntagmatiques

Les fonctions lexicales syntagmatiques modélisent les collocations, les phénomènes de combinatoire lexicale non libre qui nous intéresseront particulièrement ici. Avant de présenter les fonctions lexicales syntagmatiques, il importe de définir précisément la notion de collocation retenue dans le cadre de la théorie Sens-Texte et, conséquemment, dans celui de la présente recherche.

Une collocation est une expression formée de deux unités lexicales A et B telle que A est choisie librement, en fonction de son sens 'A', alors que B est choisie en fonction de A pour exprimer un sens donné 'C' auprès de A.

De la définition précédente, l'adéquation des fonctions lexicales à la description des collocations ressort pleinement. En effet, le modèle de l'application d'une fonction permet d'exprimer la caractéristique principale des collocations, la nature essentiellement contrainte de l'association : dans une collocation, un élément (le collocatif) est choisi en fonction de l'autre (la base de la collocation) pour exprimer un sens dans une position syntaxique donnée. La base détermine le ou les collocatifs pouvant s'appliquer à elle.

Le modèle d'une fonction pour représenter formellement la collocation explique également l'opacité relative du collocatif à l'intérieur de la collocation. Nous avons vu que le sens d'une fonction lexicale — quand il s'agit d'une fonction lexicale standard — appartenait à un petit ensemble de significations de base, très proches de primitifs sémantiques ('non', 'grand', 'faire', etc.). C'est ce sens primitif que traduit, de façon souvent idiomatique, le collocatif.

Les fonctions lexicales syntagmatiques sont généralement présentées en fonction de la partie du discours et du rôle syntaxique du collocatif. Nous présentons ci-dessous quelques-unes des fonctions lexicales syntagmatiques les plus courantes : trois fonctions retournant des valeurs adjectivales ou adverbiales (**Magn**, **Ver** et **Bon**) et trois fonctions retournant des valeurs verbales (**Oper**$_i$, **Func**$_i$ et **Labor**$_{ij}$). Ce dernier groupe nous intéresse particulièrement puisque la méthode de repérage des collocations présentée ici va porter essentiellement sur les combinaisons verbales de la langue de l'informatique.

2.1.4.3.1 Collocatifs adjectivaux ou adverbiaux

Magn associe à une unité lexicale L l'ensemble des unités lexicales qui expriment auprès d'elle l'intensification, le sens général 'très', 'beaucoup', etc. Il s'agit de modificateurs quantitatifs de L.

Magn(*doute*) = *affreux, amer, cruel, horrible, profond*

Ver associe à une unité lexicale L l'ensemble des unités lexicales qui expriment auprès d'elle le sens 'tel qu'il se doit'. Il s'agit de modificateurs objectifs de L.

Ver(*doute*) = *fondé, justifié*

Bon associe à une unité lexicale L l'ensemble des unités lexicales qui expriment auprès d'elle le sens 'bon', 'bien'. Il s'agit de modificateurs subjectifs de L.

Bon(*doute*) = *raisonnable*

Les trois fonctions lexicales présentées ci-dessus sont souvent combinées avec la fonction lexicale paradigmatique **Anti**. Elles produisent alors une fonction lexicale complexe.

AntiMagn(*doute*) = *léger*
AntiVer(*doute*) = *infondé, injustifié*

2.1.4.3.2 Collocatifs verbaux

Nous terminons cette introduction au système des fonctions lexicales de la théorie Sens-Texte par la présentation du premier groupe de fonctions lexicales verbales, les fonctions **Oper**$_i$, **Func**$_i$ et **Labor**$_{ij}$, qui modélisent le phénomène des constructions à verbes supports. Les valeurs de ces trois fonctions lexicales sont des verbes sémantiquement vides ou vidés de leur sens dans le contexte d'une base nominale donnée (nécessairement un nom prédicatif). Les fonctions lexicales supports ont

une vocation purement syntaxique. Elles relient une unité nominale prédicative L, par l'intermédiaire du verbe support, à l'expression des actants sémantiques de L. Les fonctions lexicales de verbes supports se distinguent entre elles par le rôle syntaxique joué auprès du verbe par L et ses actants.

Oper$_i$ associe à une unité lexicale prédicative nominale L l'ensemble des verbes supports qui prennent le $i^{ème}$ actant de L comme sujet et prennent L comme premier complément. Nous pouvons illustrer cette fonction lexicale à l'aide d'un exemple simple :

(11) *Jean a des doutes sur la sincérité de Marie.*

Dans cette phrase, AVOIR est utilisé comme verbe support, il prend JEAN, le premier actant de DOUTE, comme sujet et DOUTE comme objet. La collocation présente dans cette phrase est formellement représentée de la façon suivante :

Oper$_1$(*doute*) = *avoir, concevoir, éprouver*

Func$_i$ associe à une unité lexicale prédicative nominale L l'ensemble des verbes supports qui prennent L comme sujet et qui prennent le $i^{ème}$ actant de L comme premier complément :

(12) *Le doute habite Jean.*

La collocation qui apparaît dans cette phrase est formalisée de la façon suivante :

Func$_1$(*doute*) = *habiter* [N = X]

Func$_0$ modélise une construction à verbe support intransitive.

(13) *Un doute existe quant à l'origine précise de l'objet.*

Func$_0$(*doute*) = *exister*

Labor$_{ij}$ associe à une unité lexicale prédicative nominale L l'ensemble des verbes supports qui prennent le $i^{ème}$ actant de L comme sujet, qui prennent le $j^{ème}$ actant de L comme premier complément et qui prennent L comme deuxième complément. Ainsi

(14) *Jean met la sincérité de Marie en doute.*

sera représenté par

Labor$_{12}$(*doute*) = *mettre* [N = Y *en* ~]

Ce bref aperçu des fonctions lexicales ne rend pas justice à la richesse et à la souplesse des moyens descriptifs qu'elles mettent entre les mains des lexicographes.

Bien qu'elles permettent de rendre compte d'un nombre considérable de collocations (qu'elles explicitent formellement les caractéristiques syntaxiques et sémantiques des collocations), les fonctions lexicales syntagmatiques ne couvrent cependant pas la totalité des phénomènes de combinatoire non libre. Il existe en effet un nombre tout aussi important de liens non standard entre les unités lexicales de la langue qui n'expriment pas un des sens universels représentés par le système des fonctions lexicales standard (cf. par exemple l'expression *doute obsédant*). Ces liens non standard seront représentés par un autre type de fonction

lexicale, les fonctions lexicales non standard, qui consistent généralement en une brève définition du sens de la combinaison (à laquelle peut être liée une fonction lexicale standard) :

qui gêne ou qui inquiète X continuellement (*doute*) = *obsédant*

Nous espérons que la présentation des fonctions lexicales aura permis de démontrer qu'elles offrent la caractérisation des propriétés syntaxiques et sémantiques des collocations que recommandait Hausmann. À bien des égards, les fonctions lexicales représentent le chaînon qui manquait à ce dernier pour une description systématique et complète des collocations d'une langue. Les fonctions lexicales sont également les mieux adaptées au traitement automatique de la langue : c'est sur elles que nous avons établi les patrons de repérage des collocations verbe + nom de l'informatique.

2.1.5 Synthèse préliminaire

Nous terminons cette revue des modèles lexicologiques pour la représentation des collocations par la liste des notions introduites tout au long de cette présentation et dont nous nous servirons dans la suite de notre travail.

Unité lexicale : unité de base du lexique associée à un signifié unique et à l'ensemble des formes correspondant aux variations flexionnelles du signifiant.

Vocable : regroupement des unités lexicales qui partagent les mêmes signifiants et une partie non triviale de leurs signifiés.

Prédicat sémantique (**unité lexicale prédicative**) : unité lexicale dénotant un fait ou une entité impliquant au moins un participant appelé *actant sémantique* — les actants sémantiques sont habituellement désignés par des variables du type X, Y, Z.

Collocation : expression combinant deux éléments de façon semi-idiomatique.

Base : élément de la collocation qui, sélectionné librement par le locuteur, retient son sens dans la collocation et la contrôle.

Collocatif : élément sélectionné dans l'entourage d'une base pour exprimer un sens et un rôle syntaxique donné.

Fonction lexicale : fonction **f** qui associe à une unité lexicale L l'ensemble des unités lexicales qui expriment, en fonction de L, un sens spécifique associé à **f**.

2.2 Modèles terminologiques et terminographiques

Cette section est consacrée aux modèles proposés en terminologie et terminographie pour la description des collocations. La terminologie qui a pour objet d'étude le vocabulaire d'un domaine de spécialité s'intéresse depuis relativement peu de temps aux collocations. On peut expliquer ce manque d'intérêt par le fait que la terminologie se propose avant tout d'étudier, à travers les formes linguistiques qui les représentent, les connaissances propres à une science ou à un domaine d'activité donné. Dans cette approche classique de la terminologie, les unités lexicales sont retenues pour leur valeur dénominative, indépendamment de leur contexte

d'utilisation, et pour leur capacité à s'insérer dans le système notionnel qui est censé refléter les connaissances propres à un domaine. On s'attache à décrire des concepts — les définitions terminologiques vont contenir des informations de nature encyclopédique — et les relations qui les lient (qui ressortent davantage de l'observation de phénomènes réels que de l'analyse sémantique).

Les terminologies composées sur ces bases privilégient donc les unités nominales et attachent relativement peu d'importance à la description de leur fonctionnement en langue (à la description de leur combinatoire). Cette approche classique à la pratique terminologique est remise en question depuis déjà plusieurs années par de nombreux chercheurs, terminologues et terminographes, qui préconisent une analyse des unités lexicales de la langue de spécialité qui s'attache davantage aux propriétés strictement linguistiques et, plus particulièrement, sémantiques de ces dernières. Les changements préconisés concernent particulièrement la modélisation du sens spécialisé et la description formelle de la combinatoire lexicale des termes. Cette prise de conscience doit beaucoup au développement de la linguistique de corpus et d'outils d'aide à l'analyse des données qu'ils renferment. Ces développements récents, qui ont permis d'observer le comportement des unités lexicales spécialisées *in situ*, dans les textes mêmes, ont également favorisé le développement de modèles de description plus en prise avec la dimension linguistique des unités lexicales spécialisées.

Dans les pages qui suivent, nous résumons les travaux les plus représentatifs de ces efforts de renouvellement de la pratique terminologique, tant en terminologie qu'en terminographie. Nous

présentons dans un premier temps quelques travaux parmi les plus représentatifs de l'approche sémantico-lexicale à la description terminologique (Frawley 1988; Binon *et al.* 2000; L'Homme et Dancette 2001; Dancette et L'Homme 2002). Ces travaux, qui remettent en question les fondements mêmes de la discipline, ont permis de révéler la réalité linguistique du terme. Nous abordons ensuite les travaux qui ont porté plus particulièrement sur les collocations (développement de méthodologies d'analyse et de description des collocations de la langue de spécialité) (Freibott et Heid 1990; Heid et Freibott 1991; Heid 1992; Béjoint et Thoiron 1992). Nous terminons cette section par une présentation de quelques-unes des entreprises terminographiques les plus achevées (Cohen 1986; Meynard 2000).

2.2.1 Approches nouvelles à la terminologie

2.2.1.1 Les propositions de Frawley (1988)

L'un des premiers plaidoyers en faveur d'un changement radical de la pratique terminologique se trouve dans un article de William Frawley, publié dans le *International Journal of Lexicography* en 1988. Dans cet article, l'auteur propose un nouveau modèle de dictionnaire pour consigner les sens spécialisés, modèle qui est entièrement basé sur le *Dictionnaire explicatif et combinatoire* (DEC) de la théorie Sens-Texte. Les modèles proposés jusqu'alors pour la langue de spécialité, qui s'inspirent largement de ceux existant pour la langue générale, sont en effet jugés inadéquats pour faire face aux développements rapides de vocabulaires

dans des domaines de plus en plus variés et répondre effectivement aux besoins de communication des spécialistes.

Pour Frawley, les dictionnaires spécialisés existants sont inadéquats parce que les définitions qu'ils contiennent ne donnent pas d'informations sur l'utilisation réelle, en discours, des unités lexicales qu'ils décrivent. Il donne l'exemple de WINZE, un terme de géologie, que le American Geological Institute définit de la façon suivante :

> "A vertical or inclined opening, or excavation, connecting two levels in a mine, differing from a raise only in construction." (Frawley 1988:199)

Les informations, de nature encyclopédique, données par cette définition permettent d'identifier correctement le référent de WINZE, mais elles ne renseignent nullement sur les propriétés syntaxiques et sémantiques spécifiques à ce terme et qui déterminent son fonctionnement dans la langue (sa combinatoire). De la définition donnée ci-dessus, il est impossible de déduire la façon d'exprimer les activités normalement associées à WINZE, s'il est par exemple possible d'augmenter, de diminuer, d'éliminer ou de causer un « *winze* ».

Le problème du manque d'informations des définitions terminologiques est avant tout pour Frawley un problème formel hérité de la lexicologie générale : la pratique exclusive de la définition par genre prochain et différences spécifiques a pour seul résultat des descriptions vagues. Les informations que de telles définitions donnent sur les unités lexicales apparentées au terme défini (*opening, raise, [to] connect* dans

l'exemple ci-dessus) ne sont jamais signalées explicitement. Elles sont de toute façon insuffisantes pour utiliser correctement le terme.

Le problème du manque d'informations contextuelles dont souffrent également les dictionnaires de langue générale est exacerbé en langue de spécialité par le fait que c'est le plus souvent à des spécialistes du domaine qu'est confiée la tâche de documenter le vocabulaire de la discipline considérée, le lexicographe n'ayant pas les connaissances nécessaires pour rédiger seul les articles du dictionnaire spécialisé. C'est ce manque d'expérience lexicographique, des experts scientifiques cette fois, qui rend les dictionnaires spécialisés si peu utiles comme outils de compréhension et d'utilisation des termes.

Pour Frawley, l'adoption des principes de rédaction des articles du DEC permet de résoudre les deux problèmes : en dotant le lexicographe, le seul expert en matière de rédaction de dictionnaires, d'outils formels de représentation du sens et de la combinatoire des termes et en conservant au spécialiste le rôle d'informateur qui aide le lexicographe à repérer et à codifier le sens spécialisé.

Le DEC offre le cadre nécessaire pour représenter la réalité linguistique du terme au moyen notamment d'une définition formelle, sous forme propositionnelle, qui identifie le terme de façon unique, et d'une description systématique des relations qu'il entretient avec les autres unités lexicales de la langue de spécialité, sous la forme de fonctions lexicales. Le DEC permet ainsi d'expliciter le vague des définitions terminologiques, notamment en ce qui concerne la mise en discours du terme. Nous donnons ci-dessous la définition remaniée par Frawley de

WINZE qui explicite la cooccurrence syntaxique, sémantique et lexicale de cette entrée. Frawley souligne que cette définition est incomplète parce que basée seulement sur la définition du *American Geological Institute* donnée ci-dessus. Elle devra être complétée à l'aide de géologues qui pourront répondre aux questions soulevées plus tôt à propos des activités typiquement associées à WINZE.

WINZE: noun

X = 1	Y = 2
N	N

Vertical connection of a mine level, X, with a mine level, Y.

Lexical functions:

$Func_1$(winze) = connect (A winze connects a level with something else.)
$Func_2$(winze) = connect (same as above, because 1 = 2)
Gener(winze) = excavation (A winze is a kind of excavation.)
$Qual_0$(winze) = inclined (A winze is inclined.)
S_{loc}(winze) = mine (A winze typically appears in a mine.)
S_1(winze) = level (A winze is involved with levels.)
S_2(winze) = level (same as above)

Figure 2-6 Entrée de WINZE en format DEC

Frawley montre de façon extrêmement convaincante et pratique qu'il est possible de construire un dictionnaire spécialisé à l'aide des principes de la lexicologie explicative et combinatoire (LEC). Il identifie clairement les avantages d'une telle représentation tant au niveau de l'élucidation du sens des unités lexicales représentées qu'au niveau de la mise en discours (relations sémantiques). La représentation des termes dans un DEC permet d'éliminer le vague des définitions terminologiques actuelles sans augmenter la taille des dictionnaires de façon disproportionnée. Une telle

représentation est également particulièrement bien adaptée à l'informatisation.

Il existe aujourd'hui deux dictionnaires spécialisés dans un format compatible avec celui préconisé par Frawley. Nous présentons ces deux dictionnaires dans les pages qui suivent.

2.2.1.2 Dictionnaire d'apprentissage du français des affaires

Le *Dictionnaire d'apprentissage du français des affaires* (DAFA) (Binon *et al.* 2000) est un des premiers dictionnaires spécialisés élaborés sous la forme préconisée par Frawley. Publié il y a quelques années, il représente une véritable innovation en matière de dictionnaires spécialisés, en recourrant à une forme de définition qui identifie clairement la structure actancielle de chaque terme représenté et en incluant une description systématique des collocations que le terme contrôle. Le DAFA décrit les termes économiques du français (3 200 termes répertoriés, lexèmes et phrasèmes) et leur traduction en cinq langues. Il contient également 11 000 expressions courantes de la langue du monde des affaires et de l'économie (collocations et expressions idiomatiques). Finalement, le dictionnaire identifie les ensembles de termes apparentés, synonymes et dérivés, à l'aide de 6 000 liens ou renvois.

Dans les pages qui suivent, nous décrivons la version électronique du DAFA, en portant une attention particulière à description des unités nominales. La version électronique du DAFA, implantée dans une base de données relationnelle, permet une représentation autonome des combinaisons lexicales des entrées, indépendamment des unités lexicales

qui les composent. Chaque expression enregistrée dans le champ des collocations ou des expressions de la base de données fait l'objet d'une relation avec chacun de ses constituants. On peut donc l'appeler dans les articles consacrés aux constituants respectifs.

Les combinaisons lexicales d'une entrée de dictionnaire apparaissent dès l'entrée de la chaîne de caractères correspondante dans le premier écran de l'interface de consultation (Accueil). Elles sont présentées à l'intérieur de ce premier écran dans deux listes séparées : la première liste donne les collocatifs verbaux de l'entrée recherchée ; la deuxième liste regroupe les combinaisons nominales et expressions idiomatiques formées à partir de l'entrée. À partir de cet écran, l'utilisateur peut consulter l'article de dictionnaire de l'entrée ou accéder directement à une des combinaisons lexicales qui lui sont associées dans la base de données. Ce mode de consultation du dictionnaire est particulièrement bien adapté à l'apprentissage de la langue, l'objectif principal du DAFA. La figure ci-dessous montre le premier écran affiché après avoir entré la suite de caractères *vente*.

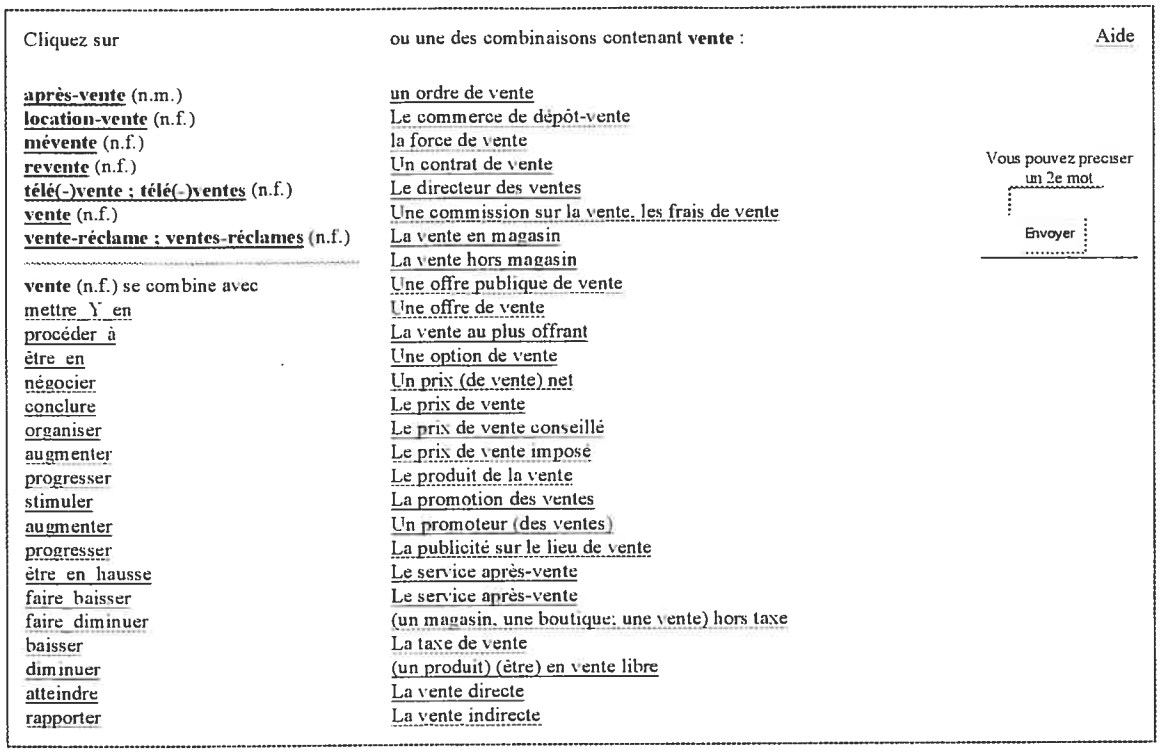
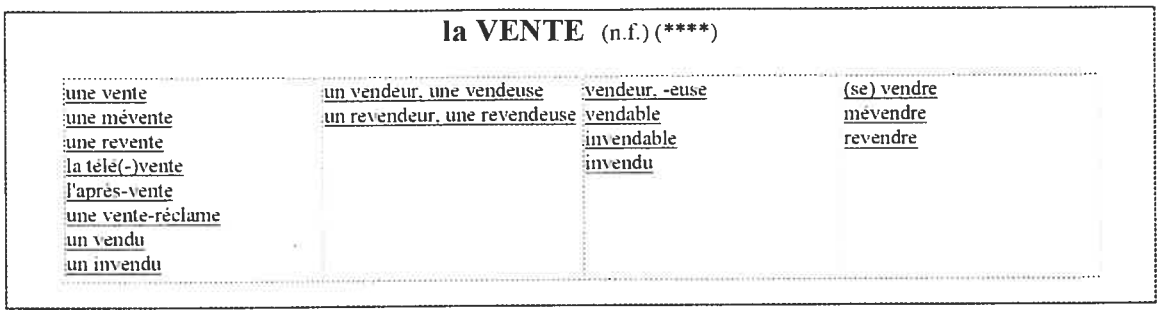


Figure 2-7 Résultats affichés pour la suite de caractères *vente*

L'article d'une entrée du DAFA donne en premier lieu la liste de ses variantes morphologiques (la *famille de mots* d'une entrée selon la terminologie du DAFA). Cette liste est présentée sous la forme d'un tableau à quatre colonnes qui organise la *famille de mots* d'une entrée selon des critères sémantiques et syntaxiques : noms dénotant des choses, noms dénotant des personnes, adjectifs et verbes.

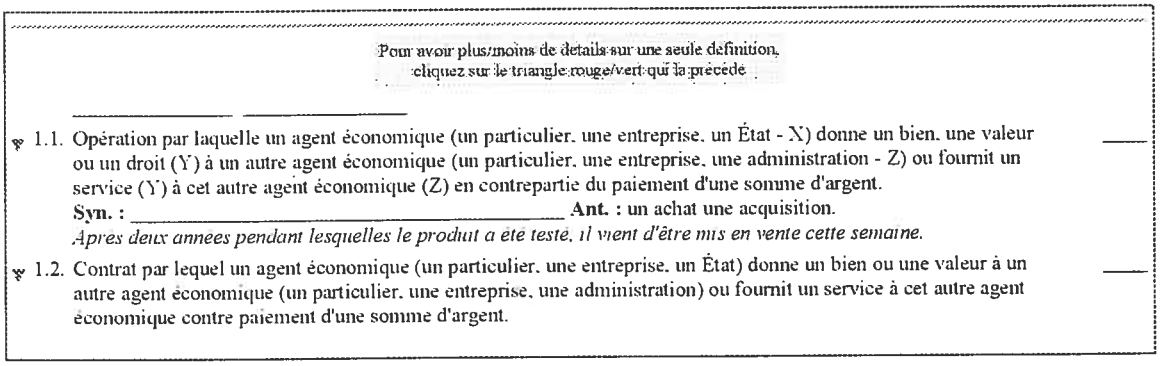


la VENTE (n.f.) (****)

une vente	un vendeur, une vendeuse	vendeur, -euse	(se) vendre
une mévente	un revendeur, une revendeuse	vendable	mévendre
une revente		invendable	revendre
la télé(-)vente		invendu	
l'après-vente			
une vente-réclame			
un vendu			
un invendu			

Figure 2-8 *Famille de mots* de VENTE

Sous ce premier tableau se trouvent les définitions des différentes acceptions de l'entrée. Les définitions du DAFA sont formulées sous une forme propositionnelle qui utilise des variables du type X, Y, Z pour représenter les différents actants sémantiques d'un terme. Elles incluent également les synonymes et antonymes éventuels de chaque terme défini et le renvoi vers ses équivalents dans chacune des cinq langues cibles.



Pour avoir plus/moins de détails sur une seule définition, cliquez sur le triangle rouge/vert qui la précède

1.1. Opération par laquelle un agent économique (un particulier, une entreprise, un État - X) donne un bien, une valeur ou un droit (Y) à un autre agent économique (un particulier, une entreprise, une administration - Z) ou fournit un service (Y) à cet autre agent économique (Z) en contrepartie du paiement d'une somme d'argent.
Syn. : _____ **Ant. :** un achat une acquisition.
Après deux années pendant lesquelles le produit a été testé, il vient d'être mis en vente cette semaine.

1.2. Contrat par lequel un agent économique (un particulier, une entreprise, un État) donne un bien ou une valeur à un autre agent économique (un particulier, une entreprise, une administration) ou fournit un service à cet autre agent économique contre paiement d'une somme d'argent.

Figure 2-9 Définitions des acceptions de VENTE

L'utilisateur a ensuite accès aux collocations (ou aux expressions idiomatiques) de l'entrée du dictionnaire en cliquant sur le lien Collocations (Expressions) au bas de l'écran Définitions. Un nouvel écran apparaît qui contient toutes les collocations enregistrées pour cette entrée, regroupées selon des critères syntaxiques (type de combinaison représentée) et sémantiques (acceptions dégagées à l'intérieur de chaque

type de combinaison). Trois types de combinaison existent pour les entrées nominales : terme + nom, terme + adjectif et terme + verbe.

Les collocations verbales d'un terme sont présentées sous la forme d'un tableau qui reprend les variables de la définition propositionnelle pour représenter les actants syntaxiques du verbe collocatif. La première colonne du tableau donne le sujet, la deuxième donne le verbe ainsi que les autres compléments et la troisième donne le correspondant nominal du verbe. Les rangées du tableau définissent les collocations particulières à l'intérieur desquelles le terme figure en position de sujet ou d'objet. Des groupes de collocations (de rangées) sont également identifiés et signalés au moyen de symboles conventionnels (▲, ▼, +, -, etc.). Il s'agit de regroupements synonymiques des verbes dont le sens est explicité dans l'écran d'Aide à l'interprétation des symboles utilisés dans les tableaux de combinaisons terme + verbe : ▲ pour signifier 'hausse', 'amélioration', ▼ 'baisse', 'détérioration', + 'situation positive', - 'situation négative', etc. La figure ci-dessous montre les collocations verbales pour la première acception de VENTE (sens 1.1).

COLLOCATIONS

vente (sens 1.1.) : Opération par laquelle un agent économique (un particulier, une entreprise, un État - X) donne un bien, une valeur ou un droit (Y) à un autre agent économique (un particulier, une entreprise, une administration - Z) ou fournit un service (Y) à cet autre agent économique (Z) en contrepartie du paiement d'une somme d'argent.

vente + **verbe** :

X	✓	mettre Y en vente procéder à une vente/à la vente de Y	la mise en vente de Y
	∀		
Y	✓	être en vente être en vente libre (expressions	- -
X et Z		négocier la vente de Y	une négociation sur la vente de Y
	∀		
X et Z		conclure la vente de Y	la conclusion de la vente de Y
X		organiser la vente de Y	l'organisation de la vente de Y
une mesure	△	(faire) augmenter les ventes (de Y)	une augmentation des ventes (de Y)
		(faire) progresser les ventes (de Y)	une progression des ventes (de Y)
		stimuler les ventes (de Y)	une stimulation des ventes (de Y)
→les ventes (de Y)		augmenter progresser	une augmentation des ventes (de Y) une progression des ventes (de Y) une hausse des ventes (de Y)

Figure 2-10 Tableau des combinaisons terme + verbe de VENTE

Le DAFA se présente avant tout comme un dictionnaire d'apprentissage. Comme les dictionnaires d'apprentissage vus précédemment (*OALDCE*, *LDOCE*, *BBI*), il utilise le champ des collocations de la base de données relationnelle pour décrire tous les phénomènes de combinatoire non prévisibles, la cooccurrence lexicale mais aussi grammaticale (le régime des unités prédicatives).

Le manque de systématicité dans les définitions terminologiques peut dérouter l'utilisateur éventuel de ce dictionnaire : une seule acception de l'entrée de dictionnaire est généralement représentée sous forme propositionnelle (cf. l'écran Définitions des acceptions de VENTE dans la Figure 2-9 ci-dessus).

Le DAFA n'en reste pas moins une ressource terminologique exceptionnelle par la richesse des informations fournies au sujet de chaque terme, particulièrement en ce qui concerne sa combinatoire. Les collocations représentées sont nombreuses et variées (cf. pour le seul terme VENTE les collocations *mettre en vente*, *être en vente*, *négocier*, *organiser une vente*, *les ventes atteignent*, *rapportent*, etc.).

Le DAFA se distingue également des autres dictionnaires spécialisés par l'utilisation de moyens formels (définitions sous forme propositionnelle, utilisation de variables) pour représenter le sens des unités terminologiques ainsi que le sens des collocations verbales d'un terme (les symboles conventionnels associés à certains regroupements synonymiques de combinaisons verbales spécialisées). Le système de symboles utilisé dans le DAFA pour représenter le sens des collocatifs verbaux des termes est repris et développé plus encore par les auteurs de ce dictionnaire dans le DAFLES (*Dictionnaire d'apprentissage du français langue étrangère ou seconde*) (Verlinde *et al.* 2003).

2.2.1.3 Version informatisée du Dictionnaire analytique de la distribution

Nous décrivons maintenant le deuxième dictionnaire élaboré selon les principes de description terminologique préconisés par Frawley. Il s'agit de la version informatisée d'un dictionnaire spécialisé, le *Dictionnaire analytique de la distribution* (Dancette et Rhétoré 2000). L'informatisation du *Dictionnaire* a été réalisée au cours de l'année 2002 par une équipe dirigée par Jeanne Dancette et Marie-Claude L'Homme au sein de l'Observatoire de linguistique Sens-Texte (OLST) de l'Université de

Montréal (Dancette et L'Homme 2002). La formalisation sur support électronique de ce dictionnaire spécialisé s'inscrit dans la lignée des travaux menés par L'Homme depuis plusieurs années sur l'application des principes de la lexicologie explicative et combinatoire (LEC) à la description des terminologies en général et de la combinatoire spécialisée en particulier (L'Homme 2002; L'Homme 2004a). Ses recherches ont permis d'éclairer trois domaines fondamentaux d'application de la LEC en terminologie :

1. Analyse du sens spécialisé (en utilisant les critères formels de désambiguïsation du sens lexical de la LEC).

2. Description formelle du sens des unités terminologiques, particulièrement des unités prédicatives. L'Homme a notamment proposé un modèle de forme propositionnelle qui intègre une caractérisation conceptuelle des actants sémantiques des verbes.

3. Description formelle des relations qui existent entre les unités terminologiques d'un domaine de spécialité.

Le projet présenté ici concerne ce troisième point. Il s'agissait de formaliser la description des relations intralinguistiques, dans les articles très denses du dictionnaire original, afin de faciliter l'apprentissage de ces relations par les utilisateurs du *Dictionnaire*. Le modèle de description retenu se base sur les fonctions lexicales. Il est implémenté dans une base de données relationnelle.

Le point de départ est le *Dictionnaire analytique de la distribution*, un dictionnaire encyclopédique très spécialisé. Il contient 350 articles qui représentent les concepts clés de la distribution. À l'intérieur de chaque article, le concept est décrit de façon encyclopédique au moyen de neuf rubriques. La première rubrique donne la ou les vedettes anglaises de l'article (tous les termes utilisés pour dénommer le concept décrit dans l'article). Elles sont suivies, dans la deuxième rubrique, de leurs équivalents français. Le matériel définitoire est réparti dans les quatre rubriques suivantes, *Définition*, *Précisions sémantiques*, *Relations internotionnelles* et *Compléments d'information*. Nous reproduisons ci-dessous l'article de MAGASIN PARASITE.

1. **PARASITE STORE, INTERCEPT STORE**

2. *MAGASIN$_{nm}$ PARASITE, MAGASIN$_{nm}$ INTERCEPTEUR, MAGASIN$_{nm}$ DE FLUX*

3. Définition :

Magasin de détail dont le **pouvoir d'attraction** (PULLING POWER) est extrêmement faible et qui profite **du flux de clientèle** (TRAFFIC) créé par les magasins voisins pour son propre **achalandage** (*goodwill*).

4. Précisions sémantiques :

Le **magasin parasite** ne possède pas d'attributs particuliers susceptibles de le démarquer des autres magasins pour attirer les clients.

5. Relations internotionnelles :

Le **magasin parasite** s'oppose au **magasin de destination** (DESTINATION STORE) qui possède, par définition, un fort pouvoir d'attraction.

Le **magasin parasite** et le **magasin satellite** (*satellite store*) sont proches. Le magasin satellite est un **magasin parasite** implanté dans un centre commercial où il bénéficie du flux de circulation des **locomotives** (ANCHORS).

6. Compléments d'information :

7. Informations linguistiques :

8. Contextes :

Another type of outlet, called a parasite store, does not create its own traffic and has no real trading area of its own. The store depends on customers who are drawn into the location for other reasons. (Berman et Evans 1995 : 273)

#L'emplacement à l'intérieur d'un centre constitue un aspect important de la stratégie de marketing d'un détaillant. Le magasin doit-il être situé entre ceux qui exercent un plus grand pouvoir d'attraction ("magasin de flux" par opposition à "magasin de circulation")? (Gaudin *et al.* 1993 : 277)

9. Exemples :

-kiosque à journaux ou fleuriste situé près d'un grand magasin ou d'un hypermarché
-petites épiceries et magasins de commodité (ou "dépanneurs" au Québec)

Figure 2-11 Article de MAGASIN PARASITE

La rubrique *Définition* donne les caractéristiques sémantiques minimales du terme vedette. Les trois rubriques suivantes sont constituées de textes descriptifs qui développent cette définition minimale en

identifiant notamment les termes reliés à la vedette à l'intérieur du vocabulaire de la distribution (synonymes, quasi-synonymes et opposés). Nous reproduisons ci-dessous des exemples de ces descriptions pris dans l'article de MAGASIN PARASITE. (Les termes apparentés sont signalés au moyen de caractères gras.)

(15) *Le* ***magasin parasite*** *s'oppose au* ***magasin de destination*** *(DESTINATION STORE) ...*

(16) *Le* ***magasin parasite*** *et le* ***magasin satellite*** *(satellite store) sont proches ...*

Les articles du *Dictionnaire de la distribution* accordent donc une importance particulière à la description des relations sémantiques de base, relations de synonymie, d'antonymie, mais aussi d'hyperonymie et de méronymie. Il s'agit de relations fondamentales en terminologie. Elles identifient des ensembles finis de termes, ceux qui partagent une ou plusieurs caractéristiques sémantiques, permettent une classification hiérarchisée de ces ensembles et révèlent finalement l'organisation conceptuelle d'un domaine spécialisé. Dans le *Dictionnaire*, les relations sémantiques fondamentales sont majoritairement décrites dans la rubrique 5. des articles (*Relations internotionnelles*).

Les articles du *Dictionnaire* contiennent également des informations sur la combinatoire du terme vedette : elles se trouvent surtout dans les rubriques *Définition* et *Précisions sémantiques*. Voici un exemple pris dans la définition de MAGASIN PARASITE.

(17) *Magasin de détail ... qui* ***profite du flux de clientèle*** *(TRAFFIC) créé par les magasins voisins pour son propre achalandage (goodwill).*

Ainsi que le montrent les exemples ci-dessus, les informations sur les liens qui unissent les termes à l'intérieur d'une famille conceptuelle donnée (celle des établissements commerciaux dans nos exemples) sont disséminées dans le corps des 350 articles du *Dictionnaire* ; en tout, 4 000 termes français et anglais sont décrits. Les descriptions, en langage naturel, utilisent des formulations variées pour représenter les différents liens lexicaux. Le projet d'informatisation du *Dictionnaire* a donc consisté dans un premier temps à identifier les relations lexicales à l'intérieur de chaque article et à proposer une modélisation uniforme pour chaque relation identifiée (à lui assigner une fonction lexicale donnée).

Dans ce travail de formalisation, les auteurs ont porté une attention particulière à la conversion des relations sémantiques de base : comme nous l'avons vu, ce sont ces relations qui ont reçu le plus d'attention dans le *Dictionnaire*. Trois fonctions lexicales ont servi à la modélisation des relations sémantiques de type taxonomique (relations générique-spécifique). Il s'agit des fonctions lexicales **Gener** (= terme générique), **Syn** (= synonymie) et **Anti** (= antonymie). Voici des exemples de ces trois relations pour le terme MAGASIN PARASITE :

Gener(*magasin parasite*) = *magasin de détail*
Syn$_{\cap}$(*magasin parasite*) = *magasin satellite*
Anti(*magasin parasite*) = *magasin de destination*

Les fonctions lexicales **Mult** ('ensemble de...') et **Sing** ('unité minimale de...') ont été utilisées pour saisir une partie des relations méronymiques entre les termes du vocabulaire de la distribution. D'autres types de relations méronymiques ont été saisis à l'aide de nouvelles fonctions (**Part**[9] et **Phase**) :

Mult(*client*)	=	*clientèle*
Sing(*assortiment*)	=	*produit*
Part(*grand magasin*)	=	*rayon*

La conversion du *Dictionnaire* en format électronique a également visé la formalisation des dérivés sémantiques des termes et de leurs collocations. De façon générale, ces dernières sont moins bien représentées que les relations sémantiques de type taxonomique dans la version électronique du *Dictionnaire*. Selon les auteurs, l'absence d'une caractérisation formelle des composantes sémantiques des unités terminologiques (notamment des participants impliqués dans les situations dénotées par les vedettes) rend difficile le repérage et l'encodage des collocations qui peuvent apparaître dans le matériel illustratif des rubriques. Nous donnons ci-dessous des exemples de collocations représentées dans la version informatisée du *Dictionnaire*.

Oper$_1$(*enseigne*)	=	*arborer, porter*

---

[9] Cette nouvelle fonction lexicale a été proposée par Fontenelle (1997) pour modéliser la relation méronymique 'partie fonctionnelle de...'.

CausFunc$_0$(*enseigne*)[10] = *développer*

Fact$_3$(*magasin parasite*)[11] = *bénéficier, profiter* [*du flux de clientèle*]

La version informatisée du *Dictionnaire de la distribution* présentée par L'Homme et Dancette (2001) est le résultat d'un travail unique en son genre d'analyse et de description d'un vocabulaire de spécialité : environ 2 000 paires de termes ont pu être décrites à l'aide de fonctions lexicales. Les auteurs ont prouvé qu'il était non seulement possible mais également désirable d'intégrer les éléments formels de la LEC à la description terminologique. Les fonctions lexicales permettent la description des termes aux niveaux linguistique et conceptuel nécessaires à l'élaboration des taxonomies recherchées par les modèles terminologiques traditionnels.

À notre connaissance, ce dictionnaire est le seul dictionnaire spécialisé qui représente les relations entre termes à l'aide d'outils développés dans le cadre d'une théorie générale de la langue. Un dictionnaire fondamental de l'informatique, conçu dès le départ comme un DEC, est actuellement en cours de développement (L'Homme 2004b).

---

[10] Cette fonction lexicale complexe signifie 'causer l'existence de...'.

[11] Cette fonction lexicale associe généralement une entité (artefact ou organe) à un verbe de réalisation qui se comporte au niveau syntaxique comme la fonction lexicale **Func$_i$** et signifie 'réalise le but de...'.

2.2.2 Modèles de représentation des collocations en terminologie

2.2.2.1 La base de données terminologique de Heid et Freibott

Le premier modèle proposé pour la représentation des collocations en terminologie a été développé lors de l'élaboration d'une base de données lexicographique et terminologique pour le service de traduction et de documentation de l'entreprise KRUPP. La base de données développée devait remplacer les différents outils utilisés par les traducteurs et rédacteurs techniques de cette entreprise. La multifonctionnalité linguistique envisagée pour la base de données de l'entreprise KRUPP (production de textes et traduction de et vers une des trois langues représentées) a conduit ses auteurs à repenser le modèle classique des bases de données terminologiques et à développer un modèle original qui permette notamment de traiter les collocations.

Afin d'entreprendre la construction d'une base de données terminologique devant servir à la fois d'outil de rédaction et de traduction de textes techniques, les auteurs ont eu besoin de formaliser un ensemble de phénomènes propres à la langue de spécialité et qui étaient encore assez mal représentés dans les ouvrages spécialisés et banques de terminologie existants. L'élaboration d'un outil de description terminologique devant répondre aux besoins en traduction d'une entreprise industrielle va donc être l'occasion pour les auteurs de proposer l'une des premières modélisations des collocations en langue de spécialité.

Pour modéliser les collocations à l'intérieur de la base de données, les auteurs vont s'appuyer sur les modèles descriptifs développés en lexicologie et lexicographie générale et particulièrement sur les travaux de Hausmann (1979) et de Mel'čuk *et al.* (1984, 1988, 1992, 1999). Ils proposent ainsi de définir la collocation comme

> « ...une combinaison polaire ... de deux lexèmes qui a un caractère conventionnel à l'intérieur d'un groupe linguistique. » (Heid et Freibott 1991:78)

L'adjectif *polaire* dans la définition proposée ci-dessus identifie ce qui constitue aux yeux des auteurs la caractéristique principale des collocations : le mode de composition particulier de ces expressions, selon lequel l'une des deux unités lexicales (la base) détermine les unités lexicales avec lesquelles elle peut se combiner (les collocatifs). Heid et Freibott retiennent donc la distinction établie en lexicographie générale entre les deux membres d'une collocation et adoptent la terminologie de Hausmann pour les désigner.

La deuxième caractéristique retenue souligne le caractère arbitraire (« conventionnel ») des collocations. Elles sont également décrites comme étant « lexicalisées » et devant être appréhendées comme un tout, à la différence des combinaisons composées librement, décrites quant à elles comme le produit « du hasard du texte ».

Les deux caractéristiques identifiées par les auteurs vont jouer un rôle déterminant dans la modélisation et la présentation des collocations à l'intérieur de la base de données terminologique. Parce qu'elles sont composées de façon arbitraire, pour exprimer un contenu sémantique qui

n'apparaît dans aucune des unités lexicales associées, les collocations ne peuvent être traduites littéralement, sur la base seule des unités lexicales qui les composent.

La différence de statut entre la base et le collocatif a également un impact sur la place qui doit être réservée à une collocation dans un dictionnaire de traduction. Ainsi que le soulignent Heid et Freibott, dans la traduction vers une langue étrangère (dictionnaire de thème), le traducteur accède au dictionnaire par l'entrée de la base (pour laquelle il cherche un collocatif convenable). En situation de version, le traducteur, qui traduit depuis une langue étrangère, peut ne pas se rendre compte qu'il est en présence d'une collocation. Il espèrera donc trouver la collocation aussi bien dans l'entrée du collocatif que dans celle de la base.

Les besoins de la description contrastive vont donc guider la modélisation des collocations à l'intérieur de la base de données développée par Heid et Freibott. Les collocations seront traitées de la même façon que les termes : elles feront l'objet d'une entrée séparée dans la base de données. En conférant aux collocations le même statut qu'aux termes, les auteurs de la base de données peuvent les décrire à l'aide des attributs et des relations qui existent déjà pour ces derniers. Ils peuvent ainsi relier entre elles les variantes géographiques, associer synonymes et équivalents aux collocations décrites sans qu'il soit nécessaire de définir de nouvelles relations dans la base de données.

Les collocations se différencient cependant des termes au niveau de la caractérisation morphosyntaxique, celle-ci devant nécessairement rendre compte de la partie du discours de deux éléments dans le cas des

collocations. L'attribut morphosyntaxique des collocations aura ainsi les quatre valeurs suivantes : nom + verbe, nom + adjectif, adjectif + adverbe, verbe + adverbe.

La modélisation des collocations comme objets linguistiques à part entière (qui possèdent le même statut que les termes) permet aux auteurs de résoudre les problèmes que ces expressions posent dans la description contrastive.

1. Les collocations ont leur propre description d'équivalence par une relation avec une entrée (collocation ou terme) d'une autre langue.

2. Elles sont reliées, par un système de double pointage, aux entrées des bases et des collocatifs. Le système de double pointage permet d'assurer l'accès à la collocation à partir de l'un ou l'autre de ses membres. Il permet le repérage en situation de thème ou de version. Le système de double pointage permet également de renvoyer de l'entrée d'une base à tous ses collocatifs et d'un collocatif à toutes les bases avec lesquelles il apparaît. Il permet donc une description systématique de la combinatoire lexicale d'un terme donné.

Le travail entrepris par Heid et Freibott est novateur puisqu'il s'agit de la première tentative de rendre compte formellement des collocations de la langue technique. Des exemples de collocations répertoriées dans la base de données de l'entreprise KRUPP incluent *débrancher le*

convertisseur, couper la soupape, déverrouiller les différentiels, arrêter la manœuvre de la grue, interrompre le télescopage, etc.

Bien qu'ils aient adopté, pour la description des collocations de la langue spécialisée, les modèles de description les plus complets de la langue générale (ceux de Hausmann et de Mel'čuk), les auteurs n'ont retenu de ces modèles que la caractérisation syntaxique des collocations. La base de données terminologique développée par Heid et Freibott ne fournit donc pas de caractérisation sémantique des collocations représentées. (Elle ne décrit pas non plus les propriétés sémantiques des termes qu'elle contient.) Les fonctions lexicales sont ainsi rejetées parce qu'offrant une description sémantique trop générale des collocations de la langue spécialisée. Selon Heid, les dispositifs descriptifs de la combinatoire lexicale des termes doivent être adaptés au domaine modélisé et s'appuyer notamment sur la description conceptuelle du domaine :

> « ... si l'on connaît la description conceptuelle (...) les chances sont relativement meilleures qu'en langue générale qu'on puisse « prédire » les collocations possibles. » (Heid 1992:535)

Ce point de vue sera défendu par un autre groupe de chercheurs, intéressés eux aussi à la description des collocations de la langue de spécialité, dont nous allons maintenant présenter les travaux.

2.2.2.2 Méthodologie pour un dictionnaire de collocations en langue de spécialité

Les travaux présentés dans cette section, visant le recensement et la description des collocations de la langue de spécialité, ont été entrepris par

Béjoint et Thoiron, deux chercheurs du Centre de Recherches en Terminologie et Traduction de l'Université de Lyon (C.R.T.T.). Dans un article publié en 1992, les auteurs vont développer à leur tour une méthodologie de repérage et d'encodage des collocations de la langue de spécialité.

Afin de définir les collocations de la langue spécialisée et de dégager des critères de sélection correspondants, Béjoint et Thoiron vont adapter la définition de Benson (1989) et proposer de caractériser les collocations aux trois niveaux suivants :

1. Syntaxique : les collocations sont des expressions linguistiques (des affinités construites au moyen d'une syntaxe).

2. Sémantique : elles manifestent une certaine fixité (*pétrification*) tout en restant non idiomatiques.

3. Fréquence : ces affinités sont pré-discursives pour les locuteurs, qui les utilisent donc avec une certaine fréquence.

La fréquence constitue pour ces chercheurs le premier critère de repérage des collocations. Ils utilisent également un autre critère : le critère de spécificité. Ce critère leur permet d'écarter les associations, mêmes fréquentes, au sémantisme trop diffus pour mériter, selon eux, de figurer dans un dictionnaire. Comme chez Benson, il s'agit d'associations formées avec de verbes pouvant se combiner avec un très grand nombre de noms : *[to] make*, *[to] cause*, etc.

En ce qui concerne l'encodage des collocations identifiées, les auteurs adoptent une approche traditionnelle, en n'admettant que les termes (unités nominales) comme bases des collocations. Seuls les termes peuvent donc être choisis comme vedettes des articles consacrés aux collocations. Les auteurs adoptent également une approche classique à la présentation des collocations à l'intérieur des articles en préconisant notamment une catégorisation sémantique des collocations qui reflète l'organisation conceptuelle du domaine. Avant de décrire les collocations d'un domaine de spécialité, il faut selon eux dégager en premier la structure notionnelle de ce domaine qui sous-tend les relations syntagmatiques entre termes.

> "C'est la perception des affinités entre notions qui peut aider à la mise en évidence plus systématique des affinités syntagmatiques entre les termes." (Béjoint et Thoiron 1992:521)

Ils donnent l'exemple de l'unité lexicale TRAITEMENT, telle qu'elle est utilisée dans le domaine médical et dont ils proposent de classifier la cooccurrence verbale à l'aide des caractéristiques suivantes :

1. sa finalité : *avoir pour but*, *s'efforcer de*, etc.

2. sa composition : *comprendre*, *consister en*, *associer X à Y*, *combiner*, etc.

3. son action : *diminuer*, *amoindrir*, *restreindre*, *arrêter*, *mettre fin*, *compenser*, *récupérer*, *rétablir*, *prévenir*, *protéger*, *aider à*, *contribuer à*, etc.

L'article de Béjoint et Thoiron présenté ici constitue lui aussi un plaidoyer convaincant pour inclure les collocations dans les terminologies et propose également une méthodologie pour le faire. Les auteurs, qui s'appuient sur les modèles développés en lexicologie et lexicographie (particulièrement le modèle implémenté dans le *BBI*), soulignent également la nécessité d'une caractérisation du sens des collocatifs (sinon le dictionnaire de collocations spécialisé risque d'être peu utile en renvoyant l'utilisateur à un dictionnaire de langue générale pour vérifier le sens des collocatifs indiqués). Cette caractérisation doit être établie selon eux en fonction des particularités de la structuration conceptuelle du domaine traité.

De la description donnée ci-dessus pour TRAITEMENT et qui s'appuie selon les auteurs sur une catégorisation notionnelle du domaine médical, il n'apparaît pas comment une telle catégorisation peut éclairer le sens des collocatifs verbaux de cette unité lexicale (voir par exemple les collocatifs qui ont trait à l'action du traitement et qui dénotent tous des activités différentes). Il y a là un problème que le recours à un modèle des connaissances propres au domaine ne semble pas résoudre. Une explicitation des modalités (prévention, maintien, guérison) et participants du traitement médical (maladie, malade, personnel soignant) dans la définition de cette unité lexicale de la langue de la médecine permettrait, selon nous, une catégorisation plus significative de ces différents collocatifs, en fonction cette fois de composantes sémantiques.

2.2.3 Entreprises terminographiques

Parallèlement aux travaux de la recherche terminologique dont il vient d'être question (développement de méthodologies d'analyse et de description des collocations), des terminologues et terminographes ont entrepris de documenter les collocations de leur langue de spécialité. Nous présentons ci-dessous en détail deux ouvrages conçus spécialement pour documenter les collocations d'un domaine de spécialité. Nous devons également signaler les trois ouvrages suivants qui consacrent une part importante de la description terminologique à la combinatoire des termes représentés. Il s'agit du *Essential Lexicon in Accounting* de Philippe Caignon (2000), du *Vocabulaire combinatoire de la CFAO mécanique* de Claude Lainé (1993) et du *Vocabulaire des systèmes dynamiques et de l'imagerie fractale* de Silvia Pavel et Monique Boileau (édition remaniée en 2003). Ces deux derniers ouvrages, publiés par le Bureau de la traduction du gouvernement du Canada, sont également intégrés à la base de données terminologiques TERMIUM Plus®, seule base de données commerciale à contenir également des collocations.

2.2.3.1 Lexique de cooccurrents – Bourse et conjoncture économique

Le premier dictionnaire consacré entièrement aux collocations d'une langue de spécialité, le *Lexique de cooccurrents – Bourse et conjoncture économique* de B. Cohen a été publié en 1986. Élaboré par son auteur comme travail de maîtrise à l'Université de Montréal, il est devenu un ouvrage de référence, cité dans la majorité des travaux qui traitent de combinatoire spécialisée.

Le *Lexique de cooccurrents* décrit les collocations les plus usitées du domaine économique et boursier. Pour arrêter la liste de ces combinaisons, Cohen a dépouillé à la main les journaux et périodiques, tant canadiens que français, de ce secteur d'activité. Les combinaisons ont été retenues en fonction de leur fréquence, l'auteur inscrivant une marque chaque fois qu'une expression revenait. Pour Cohen, comme pour la majorité des chercheurs, linguistes et lexicographes, qui se sont intéressés à leur description, les collocations se caractérisent par une certaine fréquence en langue ; ce phénomène a été largement exploité dans les premiers programmes développés pour repérer les collocations dans les textes (Berry-Rogghe 1973; Choueka *et al.* 1983; Church et Hanks 1989). Dans le cas du *Lexique de cooccurrents*, le seuil de sélection des combinaisons a été fixé arbitrairement à vingt occurrences.

Le *Lexique de cooccurrents* organise les combinaisons retenues dans une centaine d'articles correspondant aux principaux termes du domaine économique et boursier (CROISSANCE, DÉFICIT, OFFRE, PRODUCTIVITÉ, REVENU, TITRE, VALEUR, etc.). Chaque article donne une définition précise du terme (en identifiant notamment les différents actants sémantiques) et la liste de ses cooccurrents sous forme de tableau : la partie du discours du cooccurrent est indiquée en abscisses tandis qu'en ordonnées sont représentées les classes dégagées par l'auteur pour caractériser le sens des cooccurrents retenus. Il s'agit des différentes phases du cycle économique, c'est-à-dire le début, la croissance, le déclin et la fin. Cohen a ajouté une classe « indéterminés » (pour accueillir les cooccurrents qui expriment une situation stationnaire) et une classe « autres cooccurrents » pour les cooccurrents qui n'entrent dans aucune des classes précédentes. Nous

donnons ci-dessous en exemple l'article de VENTE tel qu'il figure dans le *Lexique de cooccurrents*.

VENTE : Le fait, pour un agent économique, d'échanger une marchandise contre son prix, de la transmettre à un acquéreur.

	NOMS	VERBES (SUJET)	VERBES (OBJET)	ADJECTIFS
DÉBUT	décollage redressement reprise vague (de~)	décoller se redresser repartir reprendre		
CROISSANCE	accroissement augmentation courant (de~) fermeté poussée (de~) progression	s'accroître augmenter progresser	accroître augmenter développer gonfler pousser relancer stimuler	
INDÉTERMINÉS	rythme	évoluer se stabiliser stagner		
DÉCLIN	affaiblissement baisse chute diminution faiblesse fléchissement ralentissement réduction repli tassement	s'affaiblir s'affaisser baisser chuter diminuer faiblir fléchir se ralentir se raréfier se tasser	restreindre affaiblir diminuer réduire	faibles
FIN				
AUTRES * COOCCURRENTS			conclure effectuer faire procéder (à)	ferme

* Vente est toujours au pluriel, excepté pour ces termes.

Figure 2-12 Article de VENTE du *Lexique de cooccurrents*

Le *Lexique de cooccurrents* est un ouvrage intéressant tant par la qualité du matériel présenté que par l'originalité de la description qui en

est faite. Il s'agit d'une version simplifiée du modèle originellement développé par Cohen dans son travail de maîtrise. La description originale des collocations du langage économique s'appuyait en effet sur les fonctions lexicales de la théorie Sens-Texte. Cette description n'a pas été retenue dans la version publiée du *Lexique de cooccurrents*, l'auteur ayant jugé les fonctions lexicales difficilement accessibles pour les utilisateurs éventuels de son dictionnaire.

La décision d'abandonner les fonctions lexicales ne s'est pas faite sans quelque déchirement : en renonçant aux fonctions lexicales, Cohen était consciente qu'elle affaiblissait considérablement la capacité descriptive de son modèle et sa réutilisation possible dans d'autres domaines :

> "... en abandonnant les fonctions lexicales, nous devions abandonner toute tentative de précision des nuances de sens et retomber dans une classification qui, comparativement, laissait largement à désirer. Un peu comme si nous avions jeté le bébé avec l'eau du bain." (Cohen 1992:511)

2.2.3.2 Internet : répertoire bilingue de combinaisons lexicales spécialisées

Le *Répertoire bilingue de combinaisons lexicales spécialisées français-anglais* d'Isabelle Meynard est un ouvrage récent (2000) dont l'objectif est de documenter les combinaisons lexicales en usage dans un domaine en pleine expansion. La nomenclature est composée de 128 termes anglais et français. Les combinaisons lexicales répertoriées pour

chacun de ces termes ont été extraites d'un corpus de près de deux millions d'occurrences.

La fréquence est également l'un des critères retenus par Meynard pour identifier les combinaisons lexicales d'Internet. Contrairement à Cohen, l'auteur du *Répertoire* ne l'applique pas de façon absolue. Elle retient ainsi la combinaison *[to] establish a link*, apparue trois fois seulement, dans un ouvrage de référence important. Un autre critère utilisé pour identifier les combinaisons inclues dans le *Répertoire* était l'existence d'un équivalent dans la deuxième langue considérée. Les critères de sélection des combinaisons retenues sont décrits dans Meynard (1997).

Le *Répertoire* est un dictionnaire bilingue, divisé en deux parties, pour les termes français et anglais. Chaque partie contient 64 articles (termes d'Internet français ou anglais). La majorité de ces termes sont des lexèmes. L'auteur a retenu quelques phrasèmes. Comme dans l'ouvrage de Cohen, une courte définition explicite le sens du terme pour lequel les cooccurrents ont été rassemblés. La définition est suivie de l'équivalent en langue cible. Les collocations sont données dans deux colonnes, la colonne de gauche énumère les collocations dans la langue de départ, alors que la colonne de droite donne leur traduction. À l'intérieur de chaque colonne, les collocations sont organisées selon le type de combinaison lexicale. L'auteur a retenu les quatre types suivants : nom+terme, verbe+terme, terme+verbe, terme+adjectif. Nous donnons ci-dessous l'article de FILE.

FILE

Definition

Sequence of information, such as a program, a set of data used by a program or a document created by a user.

French base noun: fichier

Collocate noun + Base noun	
• Access to a file	Accès à un fichier
• Compression of a file	Compression d'un fichier
• Deletion of a file	Destruction d'un fichier
• Display of a file	Affichage d'un fichier
• Format of a file	Format d'un fichier
• Path of a file	Chemin d'accès d'un fichier
• Storage of a file	Stockage d'un fichier
• Transfer of a file	Transfert d'un fichier
Collocate verb + Base noun	
• To access a file	Accéder à un fichier
• To delete a file	Détruire un fichier
• To display a file	Afficher un fichier
• To edit a file	Éditer un fichier
• To exchange a file	Échanger un fichier
• To import a file	Importer un fichier
• To retrieve a file	Récupérer un fichier
• To store a file	Stocker un fichier
• To transfer a file	Transférer un fichier
• To transmit a file	Transmettre un fichier
Base noun + Collocate verb	
• File contains	Fichier contient
• File displays	Fichier s'affiche
Collocate adjective + Base noun	
• Executable file	Fichier exécutable
• Individual file	Fichier individuel
• Local file	Fichier local
• Master file	Fichier principal
• Remote file	Fichier distant

Figure 2-13 Article de FILE du *Répertoire bilingue de combinaisons lexicales spécialisées*

Le *Répertoire* de Meynard fait intervenir très peu de critères (formels ou de fréquence) dans le recensement des combinaisons retenues. Ceci explique en partie le manque d'exhaustivité dans la description des

collocations d'une unité terminologique donnée. Comparer par exemple les collocatifs retenus pour FILE avec ceux retenus pour DOCUMENT (donné dans la définition de FILE) : *[to] create*, *[to] open* et *[to] save* apparaissent dans l'article de DOCUMENT, mais pas dans celui de FILE. Certains verbes apparaissent en revanche dans de nombreux articles — ils se combinent avec un très grand nombre des termes représentés. Ces verbes, qui illustrent un mode de combinaison libre, ne méritent pas d'être inclus dans le *Répertoire* (considérer par exemple le verbe *display* dans l'exemple de FILE donné ci-dessus, également indiqué comme collocatif de 22 autres termes).

L'ouvrage de Meynard ne propose pas non plus de caractérisation sémantique des collocatifs retenus. Il s'agit néanmoins d'un ouvrage précieux. Bien qu'il ne décrive que 64 termes, il contient énormément de collocations, donnant une vingtaine de collocatifs verbaux en moyenne par entrée (*[to] accept password*, *[to] block connection*, *[to] clean virus*, *[to] visit Web site*, etc.). C'est le seul dictionnaire de ce genre qui documente les combinaisons lexicales d'un sous-domaine de l'informatique. Nous nous en sommes servie pour l'évaluation des combinaisons acquises par Colex.

2.2.3.3 Conclusion

La présentation de l'ouvrage d'Isabelle Meynard conclut cet état de l'art des modèles de représentation des collocations en langue de spécialité. Ainsi qu'il a été démontré au cours de cette présentation, la combinatoire des unités lexicales spécialisées est encore assez peu étudiée des terminologues formés à l'école classique de la terminologie. La description des collocations spécialisées n'en représente pas moins un

objectif primordial pour la terminologie, la discipline la mieux adaptée à leur étude. Pour faciliter cette entreprise, les terminologues peuvent s'appuyer sur des programmes qui extraient les combinaisons lexicales significatives des textes. Nous présentons maintenant quelqu'un de ces programmes.

2.3 Modèles informatiques pour l'extraction de collocations : état de l'art

2.3.1 Introduction

Cette section est consacrée à la présentation des modèles développés pour l'extraction automatique de collocations à partir de corpus électroniques. Les programmes présentés utilisent des méthodes essentiellement statistiques pour identifier les combinaisons usuelles de la langue. L'approche statistique à la description des collocations est née des travaux de l'école contextualiste, dont il a été question au tout début de ce chapitre. Selon cette école, les collocations sont essentiellement des phénomènes textuels : on ne peut les étudier que dans les textes. Les travaux du contextualisme vont donc favoriser l'utilisation de corpus de plus en plus larges et le développement de méthodes d'analyse quantitatives afin d'interpréter les données textuelles et de révéler les associations significatives d'un texte.

Les contextualistes et, après eux, les chercheurs qui sont intéressés aux propriétés distributionnelles des collocations, vont donc proposer une définition statistique de la collocation : une paire de mots-formes qui co-occurrent significativement plus qu'ils ne le feraient si leur

combinatoire était soumise aux seules lois du hasard. John Sinclair, qui a continué les travaux de Firth et de Halliday, définit la collocation de la façon suivante :

> "... the occurrence of two items in a context within a specified environment. Significant collocation is a regular collocation between two items, such that they co-occur more often than their respective frequencies and the length of the text in which they appear would predict." (Sinclair 1970 : 150)

Les premiers programmes développés ont porté sur des mots-formes apparaissant ensemble dans un espace textuel court et ont implémenté des mesures différentes pour déterminer la force de l'association entre deux cooccurrents donnés (Berry-Rogghe 1973; Choueka *et al.* 1983; Church et Hanks 1989). Le contexte d'une combinaison lexicale correspond le plus souvent à une fenêtre de ± 5 mots à partir du mot pris comme base de la collocation à chaque itération du programme, chaque mot du texte étant traité comme une base potentielle dans des calculs successifs. Les travaux qui ont suivi ont montré les limites de ces premières approches : en effet, les paires acquises sur les seules bases statistiques ne décrivent pas toutes des collocations. Elles représentent aussi d'autres types d'associations, notamment des associations entre unités lexicales sémantiquement apparentées (cf. les premiers exemples rapportés par Church et Hanks, les paires formées des mots-formes *doctors* et *nurses* ou *doctors* et *dentists*).

Pour corriger ces problèmes et améliorer la pertinence des résultats obtenus, les chercheurs ont ajouté des traitements linguistiques aux

mesures statistiques utilisées pour repérer les collocations. Dans l'un des premiers programmes développés qui combinent les approches linguistique et statistique, Smadja (1993) utilise des critères linguistiques pour évaluer la qualité des combinaisons extraites sur des bases strictement statistiques. Les approches plus récentes appliquent le traitement linguistique en premier et basent l'extraction des collocations sur des corpus enrichis de données linguistiques. Dans ces programmes, les mesures statistiques sont ensuite utilisées pour filtrer les combinaisons atypiques (Grefenstette et Teufel 1995; Lin 1998; Kilgarriff et Tugwell 2001; Goldman *et al.* 2001).

Dans les pages qui suivent, nous présentons les travaux les plus significatifs en matière d'acquisition automatique de collocations[12]. Nous discutons en premier des modèles qui utilisent les seules statistiques et nous présentons ensuite les modèles hybrides qui combinent le traitement linguistique et statistique des données textuelles.

2.3.2 Modèles statistiques

2.3.2.1 Le score Z

L'une des premières tentatives d'acquisition automatique de collocations est décrite dans l'article de Berry-Rogghe (1973). L'auteur

---

[12] Un résumé des différentes approches à l'extraction automatique des collocations qui donne également les formules utilisées se trouve dans Daille (2001).

entreprend d'expliciter la notion introduite par Firth en 1951, notion qui définit pour le fondateur du contextualisme un niveau d'analyse autonome du sens linguistique, à travers les associations régulières de mots à l'intérieur des phrases. Les progrès de l'informatique vont permettre de mesurer l'applicabilité du modèle contextualiste sur un ensemble de textes et de décrire plus précisement ce niveau d'analyse collocationnel. Firth ayant décrit la collocation en termes assez vagues, l'auteur va s'appuyer sur la formulation, plus opératoire, de Halliday :

> "The syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at *n* removes (a distance of *n* lexical items) from an item *x*, the items *a*, *b*, *c* ..." (Halliday 2002:61)

Le programme doit donc compiler la liste des cooccurrents significatifs d'une unité lexicale *L* (prise comme point de départ des combinaisons extraites). Pour être significatifs, les cooccurrents de *L* doivent se combiner à elle avec une probabilité supérieure à celle de lexies combinées au hasard. Berry-Rogghe développe le score Z qui mesure la différence entre les fréquences observées pour chaque combinaison formée à partir de *L* dans une fenêtre de mots-formes donnée et les fréquences attendues sous l'hypothèse du hasard. Plus le score d'une combinaison est élevé, plus cette combinaison est significative.

Le corpus utilisé est de 72 000 mots-formes (71 595). Il est composé de textes disponibles alors en format électronique, un roman de Charles Dickens (*A Christmas Carol*) et deux pièces contemporaines.

Berry-Rogghe évalue la méthode développée avec l'unité lexicale HOUSE (cette unité est relativement fréquente dans le corpus analysé). Le

programme dresse la liste alphabétique des cooccurrents de HOUSE dans la fenêtre sélectionnée et compte la fréquence (le nombre d'occurrences) des paires ainsi formées. Il récupère la fréquence de HOUSE et de ses cooccurrents d'une liste alphabétique qu'il a préalablement créée et mesure le score Z de chaque cooccurrent. La fenêtre adoptée pour cette première évaluation est de trois mots-formes de chaque côté de HOUSE. Elle ne considère pas les fins de phrase.

Le tableau suivant donne la liste des cooccurrents les plus significatifs de HOUSE ordonnés selon leur score Z :

SOLD	24.0500	ONLY	2.0441	DID	0.6462
COMMONS	21.2416	COULD	1.9887	ABOUT	0.5363
DECORATE	19.9000	SOMETHING	1.9026	BUT	0.3641
THIS	13.3937	UP	1.8829	NOT	0.3221
EMPTY	11.9090	HAVE	1.8682	LIKE	0.2833
BUYING	10.5970	IN	1.7299	HIS	-0.0385
PAINTING	10.5970	MYRA	1.7232	WAS	-0.0890
OPPOSITE	8.5192	OTHER	1.6889	KNOW	-0.1038
LOVES	6.4811	BEFORE	1.6451	ALL	-0.1060
OUTSIDE	5.8626	TONY	1.4459	WELL	-0.1209
LIVED	5.6067	GHOST	1.3916	FOR	-0.1794
FAMILY	4.3744	MORE	1.3740	IF	-0.2197
REMEMBER	3.9425	MUCH	1.3227	IT	-0.4368
FULL	3.8209	WHERE	1.2896	THEY	-0.5175
MY	3.6780	ONE	1.2879	YES	-0.5818
INTO	3.5792	GET	1.1949	BE	-0.6557
THE	3.2978	OUT	1.1348	I	-0.6865
HAS	2.9359	OR	0.9316	DO	-0.6993
'RE	2.5999	PEOPLE	0.9220	WITH	-0.9090
NICE	2.3908	OF	0.9096	TO	-1.6660
YEARS	2.3712	MOTHER	0.8558	THAT	-1.8030
IS	2.1721	SEE	0.8503	YOU	-2.6034
EVERY	2.0736	BEEN	0.7713	AND	-2.6488

Tableau 2-I Collocatifs de *house* ordonnés selon le score Z

Ainsi que l'illustre le Tableau 2-I, les mots grammaticaux figurent parmi les cooccurrents les plus significatifs de HOUSE dans le corpus

analysé par Berry-Rogghe. Afin d'améliorer la pertinence des résultats obtenus, l'auteur expérimente avec la taille de la fenêtre : elle passe de 3 à 6 mots-formes de chaque côté de HOUSE pour être finalement fixée à quatre mots-formes. Ceci lui permet d'extraire un plus grand nombre de cooccurrents significatifs. L'élargissement de la fenêtre est particulièrement bénéfique pour *A Christmas Carol*, le texte qui présente les phrases les plus longues.

Le modèle d'extraction développé par Berry-Rogghe est entièrement basé sur la proximité linéaire des mots-formes considérés comme faisant partie d'une collocation. L'auteur reconnaît le bien-fondé d'une méthode qui examinerait également les relations syntaxiques entre les différents cooccurrents extraits mais conclut qu'une telle méthode est impraticable dans l'état actuel de l'analyse des textes.

Elle propose cependant d'améliorer l'extraction en excluant systématiquement les mots grammaticaux comme cooccurrents de *L*. Cette stratégie permet également d'augmenter la taille de la fenêtre (elle passe de 4 à 6 ou 7 mots-formes selon les cas) et de récupérer des cooccurrents plus éloignés de *L*.

2.3.2.2 Acquisition automatique d'expressions idiomatiques et semi-idiomatiques à partir d'un grand corpus

Dans l'une des premières expériences d'acquisition automatique à partir de grands corpus, Choueka *et al.* (1983) présentent une méthode pour extraire les expressions caractéristiques d'un texte. Les auteurs offrent la définition suivante des expressions recherchées :

> "... a sequence of two or more *consecutive* words that constitutes an autonomous linguistic unit and has acquired, because of recurrent use in specialized contexts, a meaning or a connotation that somehow transcend the ordinary meaning of its constituents." (Choueka *et al.* 1983 : 34)

La définition portant sur des mots-formes consécutifs, elle décrit surtout des expressions entièrement idiomatiques et quasi-idiomatiques (c'est-à-dire des locutions), en grande partie nominales. Parmi les exemples d'expressions que donnent les auteurs, citons *once upon a time, by and large, time and again, research and development, Merry Christmas, hit and run, United Nations, Security Council,* etc. Ces expressions constituent autant d'entrées autonomes. Elles correspondent à un certain type de stockage pour les locuteurs qui peuvent habituellement les compléter lorsqu'ils en entendent le premier mot.

L'approche adoptée par les auteurs reposant seulement sur l'analyse statistique de la distribution des mots-formes dans le texte, elle va nécessiter une quantité importante de textes. Pour l'expérience décrite dans l'article de 1983, les auteurs ont utilisé une base de données contenant près de 200 volumes de décisions rendues par les tribunaux rabbiniques, représentant plus d'un siècle de procédure judiciaire et provenant de 20 pays différents. En tout, la base de données comprend 37 500 textes rédigés en hébreu, représentant un total de 38 millions d'occurrences.

La première approche utilisée par Choueka *et al.* pour repérer les expressions usuelles de deux mots dans la base de données de textes

rabbiniques consiste à relever les paires de mots-formes consécutifs les plus fréquentes. Cette approche ne donne pas les résultats escomptés : les paires les plus fréquentes ne décrivent pas des expressions idiomatiques mais plutôt des combinaisons accidentelles entre les mots les plus fréquents de la base de données.

Les auteurs cherchent alors à mesurer la régularité de l'association d'un mot *w* avec son cooccurrent le plus fréquent et développent à cet effet un indice de sélectivité (*neighbour-selectivity index* ou *NSI*) qui tient compte également du nombre et de la fréquence locale des autres cooccurrents de *w* (leur fréquence en tant que co-occurrents de *w*).

Afin d'évaluer la mesure développée pour repérer les expressions caractéristiques d'un texte, les auteurs ont calculé le NSI des mots-formes les plus fréquents de la base de données (le rapport que ces mots entretiennent avec leur cooccurrent le plus fréquent). Ils ont ensuite ordonné les mots-formes selon leur score NSI et ont fait évaluer les 300 premiers mots par des spécialistes du domaine. On demandait aux spécialistes de deviner le cooccurrent usuel de chaque mot figurant sur la liste sachant que le mot pouvait former le début d'une expression idiomatique. Les auteurs ont ensuite comparé les réponses données par les spécialistes aux scores NSI des 300 mots-formes évalués et ont trouvé une forte corrélation entre le score NSI d'un mot et sa participation à la formation d'une expression idiomatique.

L'expérience décrite dans l'article de Choueka *et al.* présente des résultats impressionnants mais néanmoins limités puisque l'indice de sélectivité développé ne s'applique qu'à des suites de mots consécutifs.

2.3.2.3 Recours à l'information mutuelle

Dans une autre expérience déterminante pour l'acquisition automatique de collocations, Church et Hanks (1989) développent une mesure basée sur la notion d'information mutuelle de la théorie de l'information pour calculer la force de l'association entre deux unités lexicales. L'information mutuelle compare la probabilité d'observer deux mots-formes, *x* et *y*, ensemble (probabilité de la dépendance) avec la probabilité d'observer *x* et *y* séparément (probabilité de l'indépendance). Si une véritable relation lexicale existe entre *x* et *y*, la probabilité de la dépendance sera beaucoup plus élevée que la probabilité de l'indépendance et l'information mutuelle de la paire (le rapport des deux probabilités) sera largement supérieure à zéro. La paire sera alors retenue comme étant significative.

Church et Hanks vont mesurer l'information mutuelle de mots-formes qui co-occurrent à l'intérieur d'une fenêtre de 5 mots. La fenêtre choisie pour le calcul de l'information mutuelle encode l'ordre linéaire des co-occurrents, puisqu'elle représente le contexte d'apparition de *y* (de 1 à 5 mots après *x*).

Bien qu'elle représente une nette amélioration sur l'indice développé par Choueka *et al.* — elle permet de repérer des suites de mots-formes consécutifs et non consécutifs — l'information mutuelle présente de nombreux problèmes. En effet, les combinaisons qu'elle identifie décrivent un ensemble de phénomènes difficilement utilisables en l'absence de description formelle. Il s'agit par exemple d'unités lexicales appartenant au même champ sémantique (cf. les exemples donnés par Church et Hanks

des cinq cooccurrents les plus significatifs de *doctor* : *dentists*, *nurses*, *treating*, *treat* et *hospitals*). La mesure permet également d'identifier les verbes à particules *([to] set off)*, les prépositions fortement régies (*[to] allude to, [to] adhere to, [to] amount to*, etc.) et d'autres types de combinaisons lexicales (*[to] save forests, lives, jobs, money*, etc.) qui peuvent aider le lexicographe dans l'analyse des concordances des unités lexicales les plus fréquentes d'une langue (ces concordances pouvant représenter plusieurs milliers de lignes). Church et Hanks reconnaissent eux-mêmes les limites de l'information mutuelle et proposent d'améliorer l'utilité des associations lexicales identifiées en intégrant un traitement linguistique préalable des données. Cette proposition sera reprise dans les programmes d'acquisition développés à la suite des travaux précurseurs de Church et Hanks, que nous allons maintenant examiner.

2.3.3 Modèles hybrides

2.3.3.1 Xtract

Xtract de Frank Smadja (1993) est l'un des premiers programmes hybrides développés pour l'acquisition des collocations. D'après son auteur, le programme s'inscrit dans la lignée des travaux de Choueka *et al.* (1983). Il a été développé à la même époque que celui de Church et Hanks (1989).

L'outil développé par Smadja présente de nombreuses améliorations par rapport à ces travaux antérieurs. Il élimine le problème de l'acquisition de paires de mots-formes non apparentés en tenant compte de la position relative des cooccurrents à l'intérieur de la fenêtre : seuls les cooccurrents

présentant un patron marqué de cooccurrence (qui se trouvent dans une position spécifique par rapport au nœud) seront retenus. Smadja a également généralisé la méthode d'acquisition aux combinaisons formées de plus de deux mots. La dernière contribution originale du travail de Smadja concerne l'utilisation d'un filtre linguistique pour éliminer les combinaisons qui ne représentent pas des collocations (des relations syntaxiques).

Xtract a été développé à l'aide d'un corpus étiqueté de 10 millions de mots constitué des articles boursiers du service de nouvelles de l'Associated Press. Le programme permet l'acquisition de trois types de combinaisons lexicales. Les relations syntaxiques (*predicative relations*) décrivent les collocations du discours boursier ; elles sont définies de la manière suivante :

> "A predicative relation consists of two words repeatedly used together in a similar syntactic relation." (Smadja 1993:148).

Le deuxième type de combinaisons acquises représente les phrasèmes nominaux les plus courants de ce domaine tels que *stock market* ou *foreign exchange*. Finalement le troisième type de combinaisons acquises décrit un patron syntaxique (*phrasal template*) du domaine de la Bourse. Il s'agit d'une phrase idiomatique qui associe éléments grammaticaux et lexicaux, par exemple *The Dow Jones average of 30 industrials fell *NUMBER* points to *NUMBER**.

Xtract combine les méthodes statistique et symbolique d'acquisition des collocations et opère en trois étapes. La première étape identifie les combinaisons lexicales de deux unités (bigram) en utilisant les méthodes

statistiques traditionnelles. Aucune information sur la structure de la phrase n'étant disponible à ce stade, le système utilise le principe suivant, basé sur les observations de l'auteur : la majorité des relations syntaxiques relient des lexies séparées par seulement cinq mots ; Xtract considère deux mots comme étant cooccurrents s'ils se trouvent dans la même phrase et s'ils sont séparés par moins de cinq mots.

Le programme est exécuté sur les concordances d'une unité lexicale w. Il examine à tour de rôle chacun des cooccurrents w_i apparaissant dans une fenêtre de ± 5 mots autour de w et construit, pour chaque w_i, une structure de données contenant la catégorie grammaticale de w_i (PP), la fréquence de cooccurrence de w_i avec w ($freq_i$) et la distribution de $freq_i$ dans les dix positions possibles autour de w (l'histogramme de w_i). À la fin de la première étape d'exécution de Xtract, une paire de mots-formes w et w_i est retenue si

1. la fréquence de cooccurrence de w et w_i dépasse un certain seuil — $freq_i$ est à au moins un écart-type au-dessus de la moyenne et

2. parmi les positions occupées par w_i, dans la fenêtre des cinq mots qui précèdent et qui suivent w, il y a au moins une position où w_i apparaît très fréquemment — l'histogramme de w_i contient au moins une position marquée ; un dernier critère permet d'extraire cette position de l'histogramme de w_i.

Selon Smadja, la position identifiée par le critère 2 est caractéristique de la relation syntaxique qui existe alors entre les deux

éléments d'une paire. Le critère 2 permet d'éliminer les combinaisons de mots sémantiquement apparentés telles que *doctors-dentists*, *doctors-nurses* et *doctors-hospitals* retenues par le programme de Church et Hanks.

La première étape ne permettant d'acquérir que des paires de mots, Xtract contient deux étapes supplémentaires. Dans la deuxième étape, Xtract recherche les combinaisons de plus de deux éléments ou n-grams. Il s'agit des deux autres types de combinaisons identifiés par l'auteur, les phrasèmes nominaux tels que *New York Stock Exchange* et les phrases prototypiques d'un domaine de discours particulier. Dans la troisième étape, le programme analyse les combinaisons identifiées dans la première étape et propose une description de la relation syntaxique entre les deux éléments de cette combinaison. Cette dernière étape permet d'éliminer toute combinaison retenue à la fin de la première étape qui ne peut être une collocation ; c'est-à-dire, lorsqu'aucune relation syntaxique ne peut être identifiée entre les deux éléments.

Pour acquérir les n-grams, Xtract part des paires de mots-formes w et w_i acquises précédemment — l'analyse ne porte que sur les phrases où les deux apparaissent — et applique le critère 1 aux mots qui apparaissent avant et après w (dans une fenêtre de ± cinq mots). Le seuil pour retenir un n-ième cooccurent w_j de w est arbitrairement fixé à 0,75 : w_j doit apparaître dans une position donnée autour de w au moins 75 % des fois. Cette étape permet de remplacer les combinaisons incomplètes de la première étape (*blue-stocks*) avec des combinaisons complètes (*blue chip stocks*).

Dans la troisième étape, Xtract utilise un analyseur pour étiqueter la relation entre les deux éléments de la combinaison retenue à la fin de la première étape. Il y a quatre relations retenues : verbe-objet, sujet-verbe, modificateur-nom et complément-nom. De nouveau, seules les phrases contenant les deux éléments sont analysées. Si l'analyseur ne peut étiqueter une combinaison donnée, la combinaison est rejetée. Si l'analyseur assigne plusieurs étiquettes à une combinaison donnée (par exemple sujet-verbe et verbe-objet), le programme retient l'étiquette la plus fréquente.

Afin d'évaluer les performances de Xtract, l'auteur a comparé l'analyse, par la troisième étape du programme, d'un ensemble de combinaisons — il s'agit des paires de deux éléments retenues à la fin de la première étape — avec l'analyse de ce même ensemble par un lexicographe. Il s'agissait d'un ensemble de 4 000 combinaisons sélectionnées au hasard après la première étape de Xtract. Le lexicographe devait identifier les combinaisons qu'il retiendrait pour un dictionnaire spécialisé et celles qu'il rejetterait. À la fin de l'évaluation, le lexicographe avait identifié trois catégories de combinaisons, représentées par les étiquettes YY, Y et N. Les étiquettes YY et Y identifiaient les bonnes combinaisons et N identifiait les mauvaises combinaisons. Les combinaisons Y étaient meilleures que les combinaisons YY. Prises ensemble, ces combinaisons ne représentaient que 40 % des combinaisons extraites par la première étape de Xtract. La troisième étape du programme filtre de la même façon 60 % des combinaisons retenues à la fin de la première étape : elle ne peut leur assigner une des quatre étiquettes considérées.

Smadja compare ensuite les 40 % restant aux combinaisons YY et Y retenues par le lexicographe. Il calcule ainsi le rappel de la troisième étape de Xtract (pourcentage de combinaisons retenues par cette étape par rapport à l'ensemble des combinaisons YY et Y retenues par le lexicographe). Le rappel de la troisième étape de Xtract est de 94 %. Smadja mesure ensuite la précision du programme (pourcentage de combinaisons YY et Y parmi les combinaisons retenues par la dernière étape de Xtract). Le programme extrait les combinaisons usuelles du domaine boursier avec une précision de 80 %.

Le programme développé par Frank Smadja est l'un des premiers programmes spécialisés à l'acquisition des collocations d'un discours spécialisé. Les résultats de l'évaluation présentés ci-dessus démontrent également qu'il s'agit d'un programme extrêmement performant. Il possède toutefois certaines limites.

La première limite est également un problème pour les deux programmes vus précédemment. Xtract n'identifie pas certaines collocations peu fréquentes — lorsque l'unité traitée n'apparaît pas souvent dans le corpus ou lorsque le corpus est trop petit, la distribution de ses cooccurrents n'étant alors pas assez grande.

La deuxième limite concerne l'analyse effectuée lors de la troisième étape du programme : Xtract n'utilise un analyseur qu'afin d'étiqueter la relation entre les deux éléments de la paire retenue à la fin de la première étape. Le système ne dispose d'aucune information sur l'organisation interne des combinaisons retenues ; par exemple, il propose, à partir de la paire *thwart-takeover* dans l'exemple

(18) *a share offer announced Sunday is designed to thwart a takeover bid by GAF Corp.*

le patron syntaxique *thwart *ART* takeover *N** (expression prototypique du discours boursier). L'analyse syntaxique de la phrase en question lui aurait permis d'éliminer cette dernière combinaison de sa liste de patrons syntaxiques.

Finalement, il ne semble pas que le système soit capable d'acquérir les relations syntaxiques (les collocations) verbe-complément prépositionnel (les expressions telles que *fall under control*, etc.), bien qu'il repère les combinaisons nom-préposition (appelées collocations grammaticales) suivantes: *comparison to*, *association with*, etc.

2.3.3.2 Extraction des constructions à verbes supports de Grefenstette et Teufel

Grefenstette et Teufel (1995) présentent une méthode de repérage automatique des verbes supports des nominalisations anglaises (*[to] make a proposal*, *[to] give an order*, etc.) qui combine l'analyse syntaxique locale des phrases et les statistiques simples.

La méthode vise à repérer le collocatif verbal le plus commun à une nominalisation donnée, le choix de ce verbe étant généralement arbitraire et présentant donc un problème difficile pour l'apprentissage des langues et le traitement automatique. Les auteurs s'intéressent aux constructions à verbe support les plus courantes, formées du verbe et de son complément d'objet direct.

L'originalité de la méthode réside dans l'étape de désambiguïsation sémantique qu'elle comporte. Il s'agit en effet de distinguer le sens premier d'une nominalisation (l'action du verbe correspondant) de ses autres acceptions (un nom d'artefact lié à cette action, cf. par exemple *painting* mais aussi *proposal*, *offer* ou *warning*). Grefenstette et Teufel s'intéressent à l'acception de base de la forme nominale, seule acception qui entre dans les constructions à verbe support qui les intéressent.

Afin de désambiguïser correctement les formes nominales (de distinguer l'acception de base des autres sens), les auteurs vont comparer l'environnement syntaxique des verbes et de leurs nominalisations dans le corpus d'étude. La nominalisation qui exprime l'action du verbe possède la même structure actancielle que celui-ci. L'identité des structures actancielles des verbes et de leurs nominalisations se manifeste dans des structures syntaxiques parallèles, les compléments du verbe apparaissant comme des compléments du nom de la forme nominale :

(19) *Vice President Salvador Laurel ... appealed* ***to President Corazon Aquino*** *to allow her ousted predecessor to die in his homeland.*

(20) *Mrs. Marcos made a public appeal* ***to President Corazon Aquino*** *to allow Marcos to return to his homeland to die.*

La méthode d'extraction va donc s'attacher dans un premier temps à caractériser la structure actancielle des verbes du corpus (nécessaire pour la désambiguïsation de la nominalisation). Les auteurs ne disposent pas d'un analyseur syntaxique robuste qui incorpore une analyse sémantique pour identifier les compléments du verbe (notamment ceux réalisés comme

des groupes prépositionnels). Ils utilisent donc une approximation de la structure actancielle des verbes et retiennent les trois premières prépositions (en termes de fréquence) qui apparaissent après le verbe (en postulant qu'il s'agit des prépositions régies par le verbe).

Le corpus utilisé est de 20 millions de mots-occurrences (dépêches de l'Associated Press de l'année 1989). La méthode utilise une paire de mots (le verbe et son équivalent nominal). Elle consiste en plusieurs étapes, les données recueillies au cours d'une étape servant pour l'étape suivante. Les différentes étapes sont présentées ci-dessous :

1. Génération des formes fléchies du verbe et du nom.

2. Extraction des phrases contenant les formes de surface générées à l'étape précédente (création d'un sous-corpus).

3. Étiquetage morphosyntaxique du sous-corpus (afin de distinguer les phrases illustrant les emplois verbaux de celles illustrant les emplois nominaux).

4. Analyse du sous-corpus (analyse syntaxique locale qui identifie les constituants syntaxiques de premier niveau, groupe nominal, verbal et prépositionnel).

5. Extraction des groupes prépositionnels qui suivent le verbe (le programme retient les trois prépositions qui apparaissent le plus fréquemment à l'intérieur de ces derniers).

6. Sélection des emplois nominaux pour lesquels l'une des trois prépositions identifiées à l'étape précédente suit la

nominalisation et extraction des verbes qui accompagnent ces nominalisations en position d'objet direct. Les collocatifs verbaux retenus sont ordonnés par fréquence d'occurrence.

7. Liste des verbes supports candidats pour chaque nominalisation.

Le tableau suivant donne les résultats obtenus pour une dizaine de nominalisations. La première colonne donne la paire verbe-nominalisation. La deuxième colonne donne les prépositions trouvés le plus fréquemment après le verbe (la fréquence est indiqué entre parenthèses) et la troisième colonne les collocatifs verbaux les plus communs de la nominalisation.

nominalization	*preps*	*most common main verbs*
offer-offer	for(116), in(100), to(98)	make(116), begin(37), launch(36)
discuss-discussion	with(127), in(85), at(54)	have(42), hold(42), begin(9)
demand-demand	for(37), in(28), of(22)	meet(58), press(34), increase(22)
propose-proposal	in(103), for(77), to(46)	make(28), reject(26), submit(19)
order-order	of(91), to(50), in(33)	issue(24), give(8), bring(7)
complain-complaint	about(183), of(155), to(91)	receive(20), file(12), have(10)
warn-warning	of(140), against(46), in(44)	issue(17), receive(5), make(4)
confirm-confirmation	in(30), of(28), to(10)	win(6), recommend(5), have(4)
assert-assertion	in(12), at(3), to(2)	make(3), repeat(1), dispute(1)
suggest-suggestion	to(60), in(57), of(27)	make(5), reject(5), offer(2)

Tableau 2-II Collocatifs verbaux les plus communs pour une dizaine de nominalisations

Ainsi que l'illustre le Tableau 2-II, la méthode développée par Greffenstette et Teufel identifie bien les verbes supports des nominalisations étudiées (*[to] make an offer* ou *[to] have a discussion*). Parce qu'elle est basée sur la seule fréquence, la méthode n'isole pas

nécessairement, parmi les verbes rencontrés dans l'environnement d'une nominalisation, le verbe support prototypique de celle-ci mais un autre collocatif verbal (cf. par exemple *[to] meet a demand*[13]). Un autre inconvénient à l'utilisation de la seule fréquence est qu'elle ne permet pas de distinguer entre collocatifs verbaux de même fréquence et d'identifier correctement le verbe support (cf. les nominalisations *proposal* et *suggestion* dont les deux collocatifs *[to] make* et *[to] reject* sont aussi fréquents).

2.3.3.3 Programme d'extraction de D. Lin

Le programme développé par Dekang Lin (1998) est l'un des premiers à baser l'extraction des collocations sur des corpus analysés. L'auteur s'est surtout intéressé à l'acquisition des quatre types de combinaisons suivants : verbe-objet, sujet-verbe, modificateur-nom et complément-nom. En développant son programme, Lin se donne pour objectif d'acquérir le plus grand nombre de combinaisons possibles. Le programme doit également extraire les combinaisons qui n'apparaissent que peu de fois dans le corpus (une limite habituelle des programmes d'acquisition de collocations).

Le corpus utilisé par Lin est composé d'articles de deux journaux, le *Wall Street Journal* (55 millions de mots-occurrences) et le *San Jose Mercury* (45 millions de mots-occurrences). Lin utilise un analyseur morphosyntaxique pour extraire de ce corpus les relations de dépendance

---

[13] Il s'agit d'un verbe de réalisation.

entre deux unités lexicales. Les relations de dépendance extraites du corpus sont représentées par des triplets de la forme $\{w_1, rel, w_2\}$ où w_1 correspond à la tête du constituant syntaxique, *rel* à une relation de dépendance particulière et w_2 au dépendant syntaxique. Les différents types de relations (*rel*) sont représentées dans le tableau ci-dessous :

Étiquette	**Relation entre**
N:det:D	nom et déterminant
N:jnab:A	nom et modificateur adjectival
N:nn:N	nom et complément nominal
V:comp1:N	verbe et objet direct
V:subj:N	verbe et sujet
V:jvab:A	verbe et modificateur adverbial

Tableau 2-III Types de relations de dépendance utilisées par Lin

Afin de minimiser l'impact des erreurs d'analyse, le programme d'acquisition de Lin

1. n'analyse pas les phrases de plus de 25 mots-occurrences ;
2. n'examine que les analyses complètes des phrases retenues ;
3. corrige automatiquement les erreurs d'étiquetage effectuées par l'analyseur morphosyntaxique (précision de la correction automatique de 95 %).

Les erreurs d'étiquetage les plus importantes concernent les homographes de parties du discours différentes. Des exemples d'homographes dans le dictionnaire utilisé incluent *job*, *class* et *cancel*. Ces erreurs sont corrigées automatiquement de la façon suivante.

Lorsque deux relations (deux analyses) ont été identifiées pour une paire d'homographes donnée, le programme examine la fréquence des triplets correspondants. Lin donne l'exemple de la combinaison *hold jobs* extraite 49 fois avec la relation verbe-objet et une seule fois avec la relation complément-nom. Lorsque le rapport des fréquences des triplets ambigus dépasse un seuil donné (cf. l'exemple donné ci-dessus), le programme ajoute la plus petite fréquence à la plus grande (corrigeant ainsi l'erreur d'étiquetage) et rétablit la fréquence du triplet corrigé à zéro.

Lorsque les deux analyses sont possibles — l'exemple donné par Lin est *draft accord* — le rapport des fréquences des triplets correspondants ne dépasse pas le seuil fixé. Dans ce cas-là, aucun changement n'est effectué.

Pour évaluer les performances du programme, Lin a comparé les triplets extraits d'un corpus manuellement annoté avec la base de données collocationnelles créée à partir du corpus du *Wall Street Journal*. Le corpus manuellement annoté est le corpus SUSANNE. Ce corpus contient les arbres syntagmatiques de 64 des 500 textes qui compose le Brown Corpus of American English et représente les quatre genres suivants : biographies, articles de journaux, articles de journaux académiques et romans.

Lin a d'abord converti les arbres syntagmatiques du corpus SUSANNE en arbres de dépendance. Il a ensuite extrait du corpus SUSANNE les quatre types de combinaisons évaluées (verbe-objet, sujet-verbe, modificateur-nom et complément-nom). Finalement, il a comparé les triplets extraits de SUSANNE de fréquence 2 ou plus à ceux figurant dans la base de données collocationnelles du *Wall Street Journal*.

Lorsque la relation identifiée pour une paire d'unités lexicales du *Wall Street Journal* était identique à celle existant pour cette paire dans le corpus SUSANNE, le triplet correspondant de la base de données collocationnelles était marqué correct. Si les relations étaient différentes, l'auteur vérifiait s'il s'agissait d'une erreur d'analyse du programme d'acquisition de collocations. Dans le cas d'une d'erreur, le triplet du *Wall Street Journal* était marqué incorrect. Il était marqué additionnel autrement.

Lin mesure ensuite le rappel et la précision du programme d'acquisition des collocations. Le rappel est défini comme le pourcentage des triplets corrects du *Wall Street Journal* par rapport à l'ensemble des triplets extraits du corpus SUSANNE. En considérant l'ensemble des textes contenus dans SUSANNE, il est de seulement 50,5 %. Si l'on considère seulement les articles de journaux du corpus SUSANNE, le rappel du programme d'acquisition de Lin, entraîné sur le Wall Street Journal, est meilleur : 65,3 %.

Quant à la précision, elle est définie comme le pourcentage des triplets corrects du Wall Street Journal par rapport à l'ensemble des triplets extraits de ce corpus (corrects, incorrects et additionnels). Elle est de 97,8 %.

L'évaluation du programme d'acquisition de collocations présentée ci-dessus porte vraiment sur la qualité de l'analyse effectuée par l'analyseur morphosyntaxique de ce programme. Bien qu'il adapte la formule de l'information mutuelle pour inclure une troisième variable (la relation existant entre les deux éléments d'une paire), Lin n'utilise pas la

mesure pour ordonner (filtrer) les combinaisons acquises. Elles sont toutes comparées aux combinaisons extraites du corpus SUSANNE.

Dans cette application, les collocations sont définies uniquement sur la base de critères syntaxiques. Lin écrit ainsi :

> "We use the term collocation to refer to a pair of words that occur in a dependency relationship (rather than the linear proximity of a pair of words)." (Pantel et Lin 2000:78)

Il ressort clairement de la définition ci-dessus que les combinaisons lexicales auxquelles Lin s'est intéressé ne décrivent pas des collocations mais des relations syntaxiques générales entre les unités lexicales de la langue. Une application possible de la base de données collocationnelles du *Wall Street Journal* consiste à identifier des ensembles de quasi-synonymes (sur la base de leurs dépendants syntaxiques) et automatiser ainsi la construction de thésaurus.

Alors que le programme présenté ici ne cherche pas à décrire davantage les combinaisons acquises, celui que nous allons maintenant présenter, Kilgarriff et Tugwell (2001), pousse un peu plus loin la description de combinaisons acquises elles aussi à partir de corpus analysés.

2.3.3.4 Word Sketch

Word Sketch, le programme développé par deux chercheurs de l'Université de Brighton, Adam Kilgarriff et David Tugwell (2001), est conçu dans une optique lexicographique, comme outil d'aide à l'analyse des

concordances d'un mot-forme donné. Il cherche à automatiser la tâche la plus difficile de ce travail, la désambiguïsation lexicale du vocable représenté, en identifiant les cooccurrents typiques des différents sens de ce vocable.

La version de Word Sketch disponible sur le Web (WASPS) offre également une interface permettant d'associer cooccurrents et acception dégagée par le lexicographe à partir du Word Sketch de départ.

Word Sketch extrait les collocations à partir d'un corpus annoté. Il s'agit du British National Corpus (BNC), un corpus de 100 million de mots, représentant une grande diversité de genres. De façon similaire au programme de Lin présenté auparavant, Word Sketch utilise un analyseur pour extraire du BNC des quintuplets de la forme {Rel, Word1, Word2, Prep, Pos} où Rel est une relation, Word1, la tête d'un syntagme, Word2, le dépendant syntaxique, Prep, une préposition ou une particule et Pos, la position de Word1 dans le corpus. Les quintuplets peuvent contenir des valeurs nulles pour Word2 et Prep.

En tout, le programme de Kilgarriff et Tugwell extrait 26 relations du BNC. Nous les présentons ci-dessous.

Les neuf premières relations sont des relations unaires qui permettent une caractérisation morphosyntaxique du nom et du verbe (bare-noun, possessed, passive, etc.).

Les sept relations suivantes sont des relations binaires entre la tête d'un syntagme et son dépendant syntaxique (les quintuplets correspondants ont des valeurs nulles pour Prep). C'est parmi ces relations

que se trouvent les collocations. Elles sont présentées dans le tableau ci-dessous :

Étiquette	**Relation entre**	**Exemple**
object	verbe et nom	*climb the bank*
subject	verbe et nom	*the bank refused*
adj-comp	verbe et adjectif	*grow certain*
noun-modifier	nom et nom	*merchant bank*
modifier	nom et adjectif	*a big bank*
and-or	nom et nom	*banks and mounds*
predicate	nom et nom	*banks are barriers*

Tableau 2-IV Relations binaires utilisées par Word Sketch

Chacune des relations du Tableau 2-IV, à l'exception de la relation and-or, possède une relation inverse, correspondant à un ordre inversé des unités lexicales à l'intérieur du quintuplet. La relation s'interprète alors comme 'est le dépendant syntaxique de'.

Deux relations binaires existent également pour les verbes à particules (*grow up*) et les gouverneurs de compléments en *-ing* (*tired of ving*). Les quintuplets correspondants ont des valeurs nulles pour Word2. Finalement, une relation ternaire (PP-comp) relie le gouverneur d'un syntagme prépositionnel à ce syntagme (*banks of the river*). La relation inverse 'est le complément de' est également représentée.

Les auteurs ont ainsi extrait 70 millions de quintuplets du BNC. La base de données de relations grammaticales constituée sert ensuite à établir le profil lexical d'un mot-forme donné. Nous présentons le profil lexical de DOUBT. (Nous avons modifié le tableau résultat pour alléger la présentation.)

doubt (n) BNC freq= 7664

~ about	944	4.2	of ~	518	-0.3	without ~	329	3.6	modifier	1338	-0.4	object of	1990	0.4	subject of	758	-0.3
validity	25	23.1	benefit	96	20.4	-	276	24.9	little	419	23.0	cast	248	32.8	assail	7	10.1
-	185	23.0	shadow	26	15.1	have	7	13.6	nagging	29	21.6	express	119	18.8	exist	26	7.5
ability	25	22.3	seed	14	13.2	know	5	10.4	grave	45	19.6	dispel	19	15.5	express	16	6.4
it	59	18.5	avoidance	20	10.5				lingering	28	18.6	raise	103	14.7	arise	21	6.4
wisdom	11	17.4	flicker	5	6.8				casting	21	17.6	have	588	12.0	remain	31	6.2
effectiveness	8	13.6	element	13	6.6				reasonable	86	16.0	harbour	11	11.7	surface	4	5.7
safety	9	13.5	shred	4	6.5				serious	99	14.9	throw	28	9.8	persist	6	5.2
identity	7	12.4	agony	4	6.3	in ~	922	0.8	niggling	11	14.0	entertain	11	9.4	surround	13	5.2
authenticity	6	12.1	moment	7	4.9	leave	84	21.9	slightest	24	13.9	sow	8	9.0	linger	4	4.7
fitness	5	11.8	feeling	7	4.3	never	28	19.0	cast	11	11.9	seem	24	8.4	creep	4	4.0
legitimacy	6	11.7	note	4	3.4	-	585	11.1	considerable	42	9.4	be	108	8.2	concern	6	3.9
efficacy	5	10.8				remain	17	9.1	gravest	5	9.1	remove	24	7.2			
feasibility	4	10.6				reader	6	9.0	raised	12	8.7	resolve	12	6.3			
value	7	10.4				now	12	7.8	raising	7	5.7	quell	4	6.1			
view	5	10.1	into ~	41	-0.1	seem	6	7.4	expressing	4	5.5	erase	4	6.0			
usefulness	4	9.8	throw	20	14.3	when	4	7.1	growing	12	4.3	voice	5	5.0			
viability	4	9.6	call	5	6.5	really	4	6.4	slight	7	4.0	overcome	7	4.5			
legality	4	9.6				still	10	6.0	remaining	8	4.0	share	13	4.4			
extent	5	9.3				look	6	5.2	severe	8	3.6	leave	27	3.9			
outcome	5	9.2				put	11	4.8				arouse	4	3.7			
process	5	8.8	for ~	77	-0.9	also	6	4.8				settle	5	3.6			

Tableau 2-V Profil lexical de DOUBT

Ainsi que l'illustre le Tableau 2-V, le profil lexical de DOUBT est composé de listes ordonnées de relations. Les listes présentent trois types d'informations : cooccurrents de DOUBT pour chaque relation acquise, nombre d'occurrences d'une combinaison particulière et score d'information mutuelle de la combinaison, ce dernier servant à ordonner les cooccurrents à l'intérieur de chaque liste.

La première liste du tableau représente la cooccurrence grammaticale de DOUBT. Nous avons indiqué seulement la préposition régie *about* suivie de ses compléments : *doubt about validity*, *ability*, *wisdom*, etc. Les deuxième et troisième listes donnent les gouverneurs de DOUBT lorsque cette unité lexicale se trouve à l'intérieur d'un groupe prépositionnel ; il s'agit de la relation PP-comp inversée. Des exemples de ces relations sont *of doubt*, complément de *benefit*, *shadow*, *seed*, etc., *in*

doubt, complément de *leave*, *remain*, *seem*, etc. Finalement, le tableau donne les autres relations grammaticales auxquelles DOUBT participe : la quatrième liste contient les modificateurs adjectivaux de DOUBT alors que les cinquième et sixième listes contiennent les verbes avec lesquels l'unité lexicale est reliée en tant qu'objet ou sujet.

Tel qu'il apparaîtra dans le chapitre suivant, le programme de Kilgarriff et Tugwell est beaucoup plus proche de la méthode développée dans le cadre de la présente recherche pour l'extraction automatique des collocations que les programmes vus auparavant : les auteurs extraient les relations grammaticales d'un corpus analysé et appliquent un test statistique pour ordonner les combinaisons extraites.

Comme avec le programme de Lin, les résultats obtenus par Kilgarriff et Tugwell sont cependant difficilement interprétables en raison du manque de formalisation dans l'identification et la présentation des phénomènes collocationnels. Mises à part les trois relations syntaxiques représentées par les étiquettes object, subject et modifier, les relations acquises, particulièrement les relations ternaires entre une unité lexicale et un groupe prépositionnel, décrivent rarement des collocations. Ainsi qu'il a déjà été signalé, les premières listes de cooccurrents donnés dans le profil lexical de DOUBT (cf. Tableau 2-V) représentent *grosso modo* la cooccurrence grammaticale de cette unité lexicale : *doubt about validity*, *ability*, *wisdom*, etc.

Les listes suivantes du tableau, dans lesquelles DOUBT figure à l'intérieur du groupe prépositionnel, sont particulièrement difficiles à interpréter, les combinaisons données illustrant différents types de

relations sémantiques. Il peut s'agir de collocations (*[to] leave in doubt* ou *[to] put in doubt*), mais aussi de dérivations sémantiques (cf. par exemple l'expression *reader in doubt* où le syntagme prépositionnel *in doubt* est l'expression du modificateur standard du premier (ou deuxième) actant de cette unité lexicale). Signalons également les expressions *open to doubt* ou *free from doubt* qui illustrent le régime de *free* et *open* respectivement.

Malgré les limites mentionnées ci-dessus, Word Sketch s'est révélé extrêmement utile dans le cadre de la construction d'un dictionnaire de l'anglais, comme les auteurs le soulignent dans l'évaluation lexicographique.

2.3.3.5 FipsCo

Nous terminons cette présentation de l'état de l'art en extraction automatique de collocations avec la présentation de FipsCo, le programme développé par Goldman *et al.* (2001) pour extraire les collocations et les expressions multilexémiques des textes.

Alors que les modèles hybrides présentés auparavant sont basés sur des analyses superficielles des textes, FipsCo utilise un analyseur syntagmatique du français, Fips, capable d'analyser les éléments extraposés, pronoms relatifs, constituants sujets des constructions passives, etc. Les auteurs défendent ainsi l'approche linguistique à l'extraction des collocations (basée sur une analyse complète des phrases) par rapport aux méthodes statistiques traditionnelles.

Le corpus utilisé est un corpus d'articles extraits du journal *Libération*, d'un million de mots-occurrences. Les auteurs basent

l'acquisition des collocations sur les relations de dépendance syntaxique entre deux unités lexicales. Elles constituent à leurs yeux un meilleur critère d'identification des combinaisons lexicales significatives que la simple proximité des unités lexicales dans le texte parce que

1. la distance entre les deux membres d'une collocation peut être importante et

2. les objets directs subissent de nombreuses transformations grammaticales (interrogation, passivisation, relativisation, etc.)

Selon les auteurs, seul un analyseur syntaxique robuste permet l'acquisition des liens de dépendance dans l'ensemble des contextes où ces liens se réalisent. L'extraction des collocations commence donc par l'analyse du corpus par Fips qui produit pour chaque phrase une structure normalisée (qui rétablit l'ordre canonique des constituants) à partir de laquelle les collocations sont extraites. Celles-ci correspondent aux types de combinaisons lexicales suivants :

Types de combinaisons	**Exemples**
nom adjectif	*marée noire*
adjective nom	*haute technologie*
nom nom	*thé citron*
nom prep nom	*part de marché*
nom verbe (sujet verbe)	*manne tomber*
verbe nom (verbe objet direct)	*caresser espoir*
verbe prep nom (verbe GP gouverné)	*vouer (à) échec*

Tableau 2-VI Types de combinaisons identifiés par FipsCo

FipsCo extrait également du corpus les phrasèmes déjà stockés dans le dictionnaire (collocations et locutions). Des exemples donnés par les auteurs incluent *donner raison*, *avoir lieu* ou *feu vert*. Un total de 170 000 cooccurrences sont ainsi extraites. Pour chacune, le programme mesure la distance entre les deux unités lexicales (exprimée en termes de mots-occurrences les séparant).

Afin d'identifier les collocations parmi les combinaisons extraites, les auteurs utilisent une mesure statistique (le coefficient de vraisemblance) pour ordonner les résultats. La précision du filtre statistique n'est cependant pas évaluée ; on ne connaît pas son efficacité à isoler les collocations à l'intérieur d'une liste de combinaisons extraites sur des critères syntaxiques. Goldman *et al.* font en effet porter l'évaluation sur la distance moyenne entre les deux membres d'une combinaison donnée. Ils calculent également le pourcentage de cooccurrences pour lesquelles la distance entre les deux unités lexicales est supérieure à cinq mots (la fenêtre utilisée par la majorité des programmes d'extraction basés sur les statistiques). La moyenne de ces pourcentages pour les cent premières combinaisons extraites est de 29,26 % : un peu plus du quart des

occurrences d'une combinaison donnée sont des cooccurrences « lointaines ». (Les unités lexicales sont séparées par plus de cinq mots.) Ces occurrences seraient ignorées par les programmes d'extraction basés sur les statistiques.

Ces données justifient pleinement l'approche adoptée par les auteurs à l'extraction des collocations. Elles confirment également le bien-fondé de la méthode développée dans notre recherche. Avant de présenter celle-ci, nous examinons un dernier outil d'extraction de collocations entièrement linguistique, le logiciel LogoTax.

2.3.3.6 LogoTax

Nous terminons cette présentation des outils développés pour l'extraction automatique de collocations avec la présentation de LogoTax (Ludewig 2001), un outil développé dans le domaine de l'enseignement et de l'apprentissage des langues sur des bases entièrement linguistiques.

LogoTax est un outil interactif qui permet d'acquérir les expressions idiomatiques et semi-idiomatiques de l'allemand dont l'utilisation correcte dans les phrases pose problème pour l'apprenant de cette langue. Le logiciel permet notamment d'inclure, avec chaque entrée créée dans la base de données phraséologiques, les indications concernant le mode d'emploi de cette expression (article, etc.) qui font souvent défaut dans les dictionnaires classiques.

LogoTax s'appuie exclusivement sur l'analyse linguistique (syntaxique) des phénomènes de combinatoire, la fréquence d'une combinaison n'étant pas jugée déterminante dans la vocation

d'apprentissage de l'outil. En effet, certaines combinaisons peu fréquentes peuvent néanmoins être intéressantes à répertorier pour l'étudiant. Inversement, des combinaisons très fréquentes, composées de façon régulière, peuvent ne lui poser aucune difficulté.

LogoTax permet la création d'entrées dans la base de données phraséologiques à partir de phrases (il s'agit de combinaisons repérées par l'étudiant pendant la lecture de textes électroniques). Il s'agit exclusivement de combinaisons verbe + nom.

La création d'une entrée phraséologique suit les étapes suivantes :

L'étudiant sélectionne (p. ex. à partir d'un site Web) une phrase contenant la combinaison qu'il veut entrer dans la base de données et copie la phrase exemple dans la fenêtre de saisie de LogoTax.

L'étudiant sélectionne les éléments de la combinaison (nom et verbe).

Le logiciel procède à l'étiquetage morphosyntaxique de la phrase exemple (l'étiquetage morphosyntaxique permet notamment de vérifier si l'étudiant a bien sélectionné les deux éléments (nom et verbe) de la combinaison.

Le logiciel procède ensuite à l'analyse syntaxique de la phrase de départ.

L'analyse syntaxique est suivie d'une évaluation phraséologique dont le but est d'extraire les caractéristiques de la nouvelle entrée. La caractéristique principale extraite de l'arbre à cette étape de la création d'une nouvelle entrée concerne la relation syntaxique entre le verbe et le

nom. Si aucune relation syntaxique n'a pu être identifiée entre le verbe et le nom choisis, l'étudiant doit corriger son choix initial d'éléments.

LogoTax recherche ensuite dans le corpus des phrases illustrant la même combinaison (concordances contenant les deux mêmes éléments).

Les phrases témoins sont analysées à leur tour et les combinaisons verbe + nom qu'elles contiennent sont évaluées afin de vérifier qu'elles illustrent bien la même relation syntaxique entre le verbe et le nom que la phrase de départ. L'évaluation phraséologique permet également d'extraire de la structure d'attributs associée à chaque nœud de l'arbre syntaxique les autres caractéristiques de la combinaison : il s'agit notamment pour le nom d'indications quant à son nombre et à sa détermination, pour le verbe, d'indications concernant la voix.

L'étudiant peut ensuite ajouter la combinaison à la base de données.

Parce qu'il repose sur des analyses élaborées des phrases d'un corpus témoin, LogoTax permet la création d'entrées phraséologiques de grande qualité contenant plus d'informations que les dictionnaires classiques : régime, détermination, modification (adjectivale), composition, etc.

LogoTax a été conçu comme un outil d'aide à l'apprentissage : la création d'une nouvelle entrée est entièrement motivée par les besoins de l'étudiant et la difficulté pressentie à utiliser correctement une combinaison particulière. Cette vocation explique également que les auteurs de l'outil n'aient pas cherché à décrire formellement les

expressions enregistrées dans la base de données collocationnelles (notamment le sens d'une combinaison particulière).

Le corpus utilisé par LogoTax est composé d'articles du journal Spiegel de l'année 1996. L'auteur envisage d'y ajouter un corpus parallèle (bilingue) aligné au niveau des constituants afin d'ajouter des traductions aux combinaisons créées.

La présentation de LogoTax conclut cet état de l'art sur les outils d'extraction de collocations et constitue une bonne introduction à notre méthode d'extraction, basée elle aussi sur des analyses syntaxiques élaborées. Avant de présenter celle-ci, nous examinons dans le prochain chapitre les étapes de préparation du corpus utilisé, dont la plus importante est l'analyse de ce corpus par le logiciel de TA Logos.

3 Outils utilisés pour la recherche : le corpus spécialisé et l'analyseur de Logos

La méthode de repérage des collocations de la langue de spécialité développée dans le cadre de cette recherche repose sur un corpus informatique spécialisé. Ce corpus a rempli plusieurs fonctions tout au long de notre travail. Premièrement, il nous a permis de dégager, par l'étude d'un échantillon représentatif de phrases, les patrons morphosyntaxiques qui forment la base de notre programme de repérage de collocations. Nous nous sommes également servie du corpus afin de valider les différentes versions de l'outil développé. Finalement, nous avons utilisé les informations extraites du corpus (les combinaisons verbe + nom) afin de mesurer l'efficacité du filtrage statistique.

Étant donné le rôle central joué par le corpus dans l'élaboration du programme de repérage des collocations, le choix d'un corpus d'étude convenable s'est révélé fondamental pour la recherche. Premièrement, les textes devaient se trouver sous format électronique (support informatique). Deuxièmement, ils devaient être représentatifs des usages dans le domaine de spécialité étudié, celui de l'informatique. Comme il s'agit d'un domaine très large, nous devions nous assurer que nous disposions d'un nombre suffisamment large de textes couvrant un grand nombre de sous-domaines de cette discipline. Finalement, les textes devaient être rédigés en anglais, le choix de cette langue étant déterminé par la méthode d'extraction elle-même : celle-ci base en effet l'extraction des collocations non pas directement sur les textes mais sur les arbres correspondant aux analyses de ces textes (les structures syntaxiques des phrases du corpus).

L'utilisation d'un analyseur syntaxique de l'anglais à ces fins (celui du logiciel de TA Logos) a donc déterminé la langue du corpus d'étude.

Pour créer le corpus d'étude, nous avons rassemblé des textes provenant de deux sources différentes. La source principale des textes du corpus d'étude est le corpus informatique de l'Observatoire de linguistique Sens-Texte (OLST) de l'Université de Montréal. La description du corpus fait l'objet de la première section de ce chapitre. Nous présentons d'abord les critères retenus pour l'élaboration et la gestion des corpus par les membres de l'OLST (section 3.1.1), puis la description détaillée du corpus d'étude (section 3.1.2). Dans la deuxième section du chapitre, nous examinons les étapes de traitement préalable du corpus : le nettoyage des documents et l'analyse syntaxique : la création des arbres syntaxiques dont seront extraites les collocations au moyen de l'analyseur de Logos.

3.1 Description du corpus

3.1.1 Critères de sélection des textes du corpus de l'OLST

La majorité des documents qui composent le corpus utilisé dans le cadre de la présente thèse nous ont été fournis par l'Observatoire de linguistique Sens-Texte de l'Université de Montréal. Le corpus informatique anglais de l'OLST a été élaboré au cours des dix dernières années sous la direction de Marie-Claude L'Homme. Il est entretenu par l'équipe de recherche Éclectik, la composante terminologique de l'OLST, dont l'une des priorités est la recherche et le développement de ressources informatiques pour la terminologie.

Le corpus informatique anglais de l'OLST représente une partie seulement des ressources informatiques mises à la disposition des chercheurs de ce groupe de recherche. En plus du corpus informatique, l'OLST dispose de corpus juridiques et médicaux (anglais et français), d'un corpus de la mécanique (français) et d'un corpus de la distribution (espagnol).

Afin de développer les ressources informatiques de l'OLST, l'équipe Éclectik a mis au point une procédure de sélection des textes basée sur l'application de différents critères de sélection (Marshman 2003). Les documents retenus sont ensuite décrits dans une base de données de gestion de corpus à l'aide d'un formulaire dont les principaux champs correspondent aux six critères de sélection utilisés. Nous les résumons ci-dessous.

Domaine de spécialité. Premier critère utilisé pour retenir un texte, il permet d'assurer la représentativité du corpus confectionné : le texte retenu doit illustrer le mieux possible le domaine que l'on souhaite étudier.

Langue. Ce critère concerne la langue de rédaction des textes retenus : il est préférable de retenir des textes originaux et non des traductions.

Niveau de spécialisation. Ce critère vise à identifier la situation de communication d'un texte. Celle-ci est généralement définie en fonction de l'auteur du texte spécialisé et des destinataires. Les auteurs de textes spécialisés reconnus dans la base de données de gestion de corpus de l'OLST sont des experts, des semi-experts ou

des enseignants. Les destinataires de ces textes, quant à eux, peuvent être des experts, des initiés, des non-initiés ou des étudiants. Les catégories d'auteurs et de destinataires utilisées dans la base de données de l'OLST sont basées sur Pearson (1998). Ces catégories permettent de définir les situations de communication suivantes (associées à quatre niveaux de spécialisation) : expert à expert (très technique), expert à initié (technique), enseignant à étudiant (didactique), expert ou semi-expert à non-initié (vulgarisation).

Type de document. Ce critère identifie le type de publication. Les options disponibles dans la base de données de l'OLST incluent : article de journal, article de revue générale, article de revue spécialisée, article de journal académique, article de vulgarisation, manuel technique, manuel de cours, notes de cours et autre ouvrage. Le type de publication, comme l'auteur et le destinataire, est généralement indicatif du niveau de spécialisation d'un texte.

Date de parution. Ce critère permet de s'assurer que les documents retenus sont de publication récente. On peut toutefois retenir les documents plus anciens datant de l'introduction d'une technologie donnée, qui sont susceptibles de contenir plus de contextes définitoires que les documents décrivant cette même technologie quelques années plus tard (Meyer et Mackintosh 1996).

Qualifications de l'auteur. Ce critère permet de garantir l'exactitude scientifique d'un texte spécialisé, un point particulièrement important pour la recherche terminographique. On

recommande d'utiliser des textes d'auteurs reconnus dans leur domaine bien qu'il soit parfois difficile d'évaluer ce dernier point (Pearson 1998).

En plus des informations décrites ci-dessus, la fiche signalétique de chaque texte composant un corpus de l'OLST donne également le nom et le format du fichier correspondant (les documents sont généralement stockés en format texte), le nombre de mots et une référence bibliographique complète.

3.1.2 Contenu textuel du corpus d'étude

Le corpus informatique de l'OLST renferme une majorité de documents provenant de l'Internet. Il est composé de 80 documents en format texte et couvre les aspects fondamentaux du domaine, c'est-à-dire l'utilisation, la création et la composition de systèmes informatiques. Les sujets abordés se répartissent *grosso modo* entre les six catégories suivantes :

- Comprendre l'informatique (*Maintaining Your Computer, How PCs Work, The ABC of computer security*)
- Matériel (*Computer Storage Media, Introduction to PC Hardware, Identifying A Motherboard*)
- Système d'exploitation (*Getting Started with Windows 2000*)
- Logiciel (*Voice Recognition Advances*)

- Programmation et réseaux (*Choosing a Network that's Right for You*)

- Internet (*Introduction to Web Technology, How E-mail Works*)

Ces exemples des sujets abordés dans les textes du corpus informatique de l'OLST permettent d'illustrer la variété des sous-domaines représentés : en plus de la catégorie générale « Comprendre l'informatique », cinq autres catégories identifient les grands axes de la discipline.

En ce qui concerne leur niveau de spécialisation, les textes du corpus informatique de l'OLST sont majoritairement destinés à un public de non-spécialistes (53 documents sur 80). Ces documents représentent des guides d'initiation aux technologies de l'information (*A Beginner's Guide to HTML, Introduction to PC Hardware, Introduction to Web Technology*), des articles provenant de sites Web de vulgarisation (www.pcmech.com, www.howstuffworks.com, www.pcnineoneone.com) et des manuels pédagogiques, riches en définitions de toutes sortes (*How Computers Work, Computers and information processing: concepts and applications*). Les textes didactiques constituent souvent la part la plus importante d'un corpus spécialisé parce qu'ils contiennent un grand nombre de termes et de contextes explicatifs (Meyer et Mackintosh 1996).

Le corpus compte également une moindre proportion de textes à l'usage de spécialistes, tels que les essais comparatifs, enquêtes ou études publiés dans un certain nombre de revues spécialisées et quelques documents de présentation technique destinés à des professionnels des technologies de l'information (*Integrating Windows NT and Enterprise*

Computer Systems, *The Deep Web: Surfacing Hidden Value*, *Computer Security White Paper*).

Les textes du corpus informatique de l'OLST sont des textes récents, la majorité ayant été rassemblés entre 2001 et 2002. Le corpus comprend également quelques ouvrages plus anciens dont un manuel de cours datant de 1992 (*Computers and information processing: concepts and applications*).

Dans sa forme actuelle, le corpus de l'OLST compte environ 540 000 occurrences. Nous avons augmenté la taille du corpus en lui ajoutant des textes similaires (13 nouveaux documents représentant un total de 70 000 occurrences). Nous voulions améliorer la représentativité du corpus d'étude, notamment dans les sous-domaines « Internet » et « Programmation et réseaux ». Les textes ajoutés proviennent tous de sites Web. Parmi ces textes figurent deux didacticiels : *Teach Yourself CGI Programming with Perl in a week* et *Essentials of Web Development.* Ils ont été saisis en 2003.

Dans sa forme définitive, le corpus d'étude totalise donc 610 000 occurrences. Le tableau ci-dessous donne le récapitulatif des documents composant le corpus utilisé dans la présente recherche. La liste complète des documents figure dans l'annexe A.

Sous-domaine	*Nbre de documents*	*Nbre de mots*
Comprendre l'informatique	23	217 656
Internet	15	116 359
Logiciel	5	8 333
Matériel	24	176 464
Programmation et réseaux	13	78 193
Système d'exploitation	7	14 028
	87	611 033

Tableau 3-I Récapitulatif des documents du corpus d'étude

Nous abordons maintenant les deux étapes de préparation du corpus de l'OLST en vue de l'extraction des collocations.

3.2 Traitement du corpus

3.2.1 Nettoyage du corpus

Ainsi qu'il a été mentionné auparavant, une grande partie des documents du corpus de l'OLST proviennent de l'Internet. Il s'agit de pages HTML sauvegardées au format texte (le format retenu par défaut pour les corpus de l'OLST). Avant d'être soumis au logiciel de TA, les documents ont fait l'objet d'un nettoyage préalable afin d'optimaliser l'analyse des phrases qu'ils contiennent. Nous avons tout d'abord éliminé la référence bibliographique qui figurait au début de chaque document. Un exemple de référence est donné ci-dessous.

```
RISLEY, D. (Page consultée le 26 octobre 2001). Build Your Own
PC, PC Mechanic. En ligne. Adresse URL:
http://www.pcmech.com/byopc/index.htm [buildingpc.txt]
```

Figure 3-1 Référence bibliographique d'un document du corpus

Nous avons ensuite éliminé des documents les marques de formatage héritées de la page Web originale qui rendaient impossible une analyse correcte des phrases par le logiciel de TA utilisé. Nous donnons ci-dessous un exemple de ces marques de formatage.

```
··The SMTP Server¶
··Whenever you send a piece of e-mail, your e-mail client interacts·¶
··with the SMTP server to handle the sending.·The SMTP server on your·¶
··host may have conversations with other SMTP servers to actually·¶
··deliver the e-mail.·¶
```

Figure 3-2 Extrait d'un document du corpus de l'OLST

Les espaces surnuméraires et marques de paragraphe représentés dans l'extrait ci-dessus sont en effet interprétés par l'analyseur syntaxique de Logos comme autant de fins de phrase, ce qui a un effet désastreux sur l'analyse de ce court paragraphe. Afin de nettoyer les documents du corpus informatique, nous avons créé une macro Visual Basic dans un document Word. Après avoir nettoyé les documents dans Word, nous les avons de nouveau sauvegardés au format texte.

Nous avons ensuite soumis les documents individuellement au logiciel de TA pour être analysés. Dans la section qui suit, nous examinons cette dernière étape de préparation du corpus.

3.2.2 Analyse du corpus

L'analyse du corpus par le système Logos (Logos Engine® de GlobalWare AG) génère les représentations à partir desquelles seront extraites les collocations de la langue de l'informatique. Il s'agit d'arbres représentant la structure syntaxique des phrases du corpus. Afin de décrire les représentations produites par l'analyseur syntaxique de Logos, nous présentons brièvement le système lui-même. Nous décrivons tout d'abord les différentes étapes du processus de traduction, en nous attachant particulièrement à la description de la méthode d'analyse suivie pour chaque phrase (section 3.2.2.1). Nous présentons ensuite SAL (*Semantico-syntactic Abstraction Language*), le système de représentation des unités lexicales qui constitue la base ontologique et structurale du système de TA (section 3.2.2.2). Dans la dernière section (3.2.2.3), nous comparons SAL avec un autre modèle de description du sens lexical, le système d'étiquetage sémantique du DiCo. SAL fournit également le modèle sur lequel nous baserons les patrons morphosyntaxiques d'acquisition des collocations. Nous lui consacrons donc les deux dernières sections de ce chapitre.

3.2.2.1 Présentation générale du système de TA Logos

L'approche à la traduction automatique implémentée dans le système Logos est représentative des systèmes de transfert (Schmid et Gdaniec 1996). Elle repose sur la mise en correspondance, dans un module de transfert, des structures syntaxiques des phrases source et cible. Dans un premier temps, le système construit une représentation

formelle de la structure de la phrase source (analyse). Cette structure est ensuite transformée en une structure cible équivalente. La structure cible est finalement donnée à un module de génération, dernière étape dans la traduction d'une phrase donnée. L'opération de transfert ayant produit l'arbre syntaxique qui correspond à la phrase cible, la génération consiste ici à calculer les formes fléchies des unités lexicales figurant sur les nœuds de cet arbre (génération morphologique).

Deux caractéristiques différencient cependant Logos des autres systèmes de transfert :

1. Les processus sémantique et syntaxique sont intégrés grâce à SAL, l'ensemble structuré de catégories sémantico-syntaxiques développé par les créateurs du système pour représenter les unités lexicales d'une langue.

2. L'analyseur comprend six modules différents, chaque module étant affecté à une tâche particulière et disposant de sa propre grammaire. Les deux premiers modules (RES1-2) prennent en charge l'étiquetage morphosyntaxique de la langue source. Les quatre modules suivants (TRAN1-4) effectuent l'analyse proprement dite et le transfert. Les quatre modules de TRAN correspondent à quatre niveaux d'analyse. Ils collaborent à l'élaboration de l'arbre source et à la construction de la structure cible correspondante. Les deux processus sont synchronisés entre eux par l'application des règles de la grammaire cible affectée à chaque module de

TRAN, règles qui exécutent le transfert des constituants identifiés par le module.

L'architecture générale du système de traduction anglais source (EN) est représentée à la Figure 3-3.

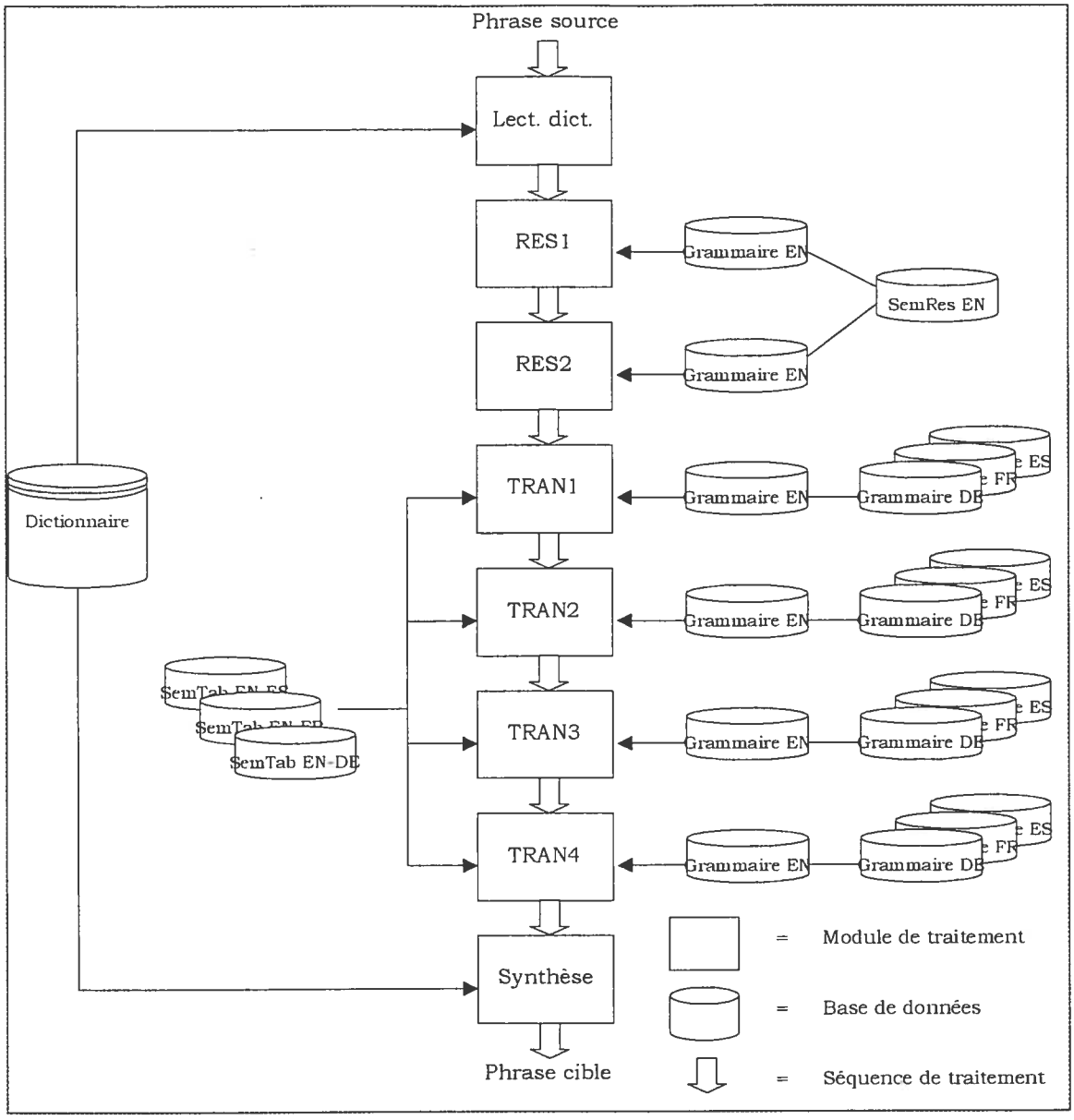


Figure 3-3 Architecture du système Logos

Ainsi que l'illustre la Figure 3-3, la traduction automatique d'une phrase commence avec la lecture du dictionnaire. Au cours de cette étape, chaque mot-forme (ou groupe de mots-formes) trouvé dans le dictionnaire est remplacé par un ensemble de valeurs qui définissent les caractéristiques sémantiques, syntaxiques et morphologiques de l'unité

lexicale associée à l'entrée de dictionnaire. (Le dictionnaire retourne plus d'un ensemble de valeurs dans le cas d'homographes.) Cet ensemble de valeurs est appelé unité sémantico-syntaxique (USS). Il représente l'unité de base du système de TA. Le format général d'une unité sémantico-syntaxique est donné à la Figure 3-4, illustré avec l'unité sémantico-syntaxique associée à PROGRAM$_N$.

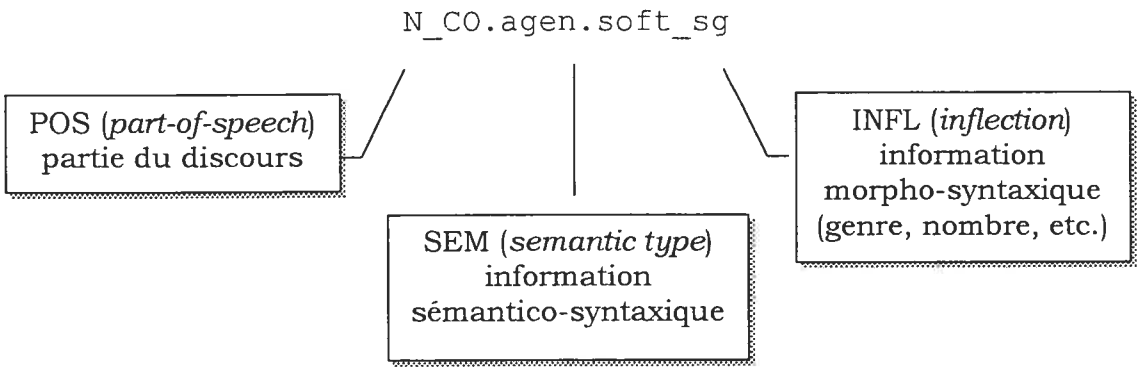


Figure 3-4 Unité sémantico-syntaxique représentant PROGRAM$_N$

Les unités sémantico-syntaxiques sont composées de trois parties. La première partie donne la partie du discours de l'unité. Celle-ci détermine les deux autres types d'information fournis : caractérisations sémantico-syntaxique et morphosyntaxique.

La partie centrale d'une unité sémantico-syntaxique représente sa caractérisation à l'un des trois niveaux de la hiérarchie SAL. Le niveau représenté dans la Figure 3-4 est le niveau le plus spécifique de la hiérarchie. Il identifie PROGRAM$_N$ comme un `software type` (`soft`), catégorie fille des `agentives` (`agen`), elle-même fille des `Concrete Nouns`. (Nous revenons plus en détail sur SAL dans la section qui suit.)

La lecture du dictionnaire a donc pour résultat de convertir la phrase source en une chaîne d'unités sémantico-syntaxiques. Le processus est illustré dans la figure Figure 3-5 avec la phrase « Other forms of storage media can also be accessed through optional devices ».

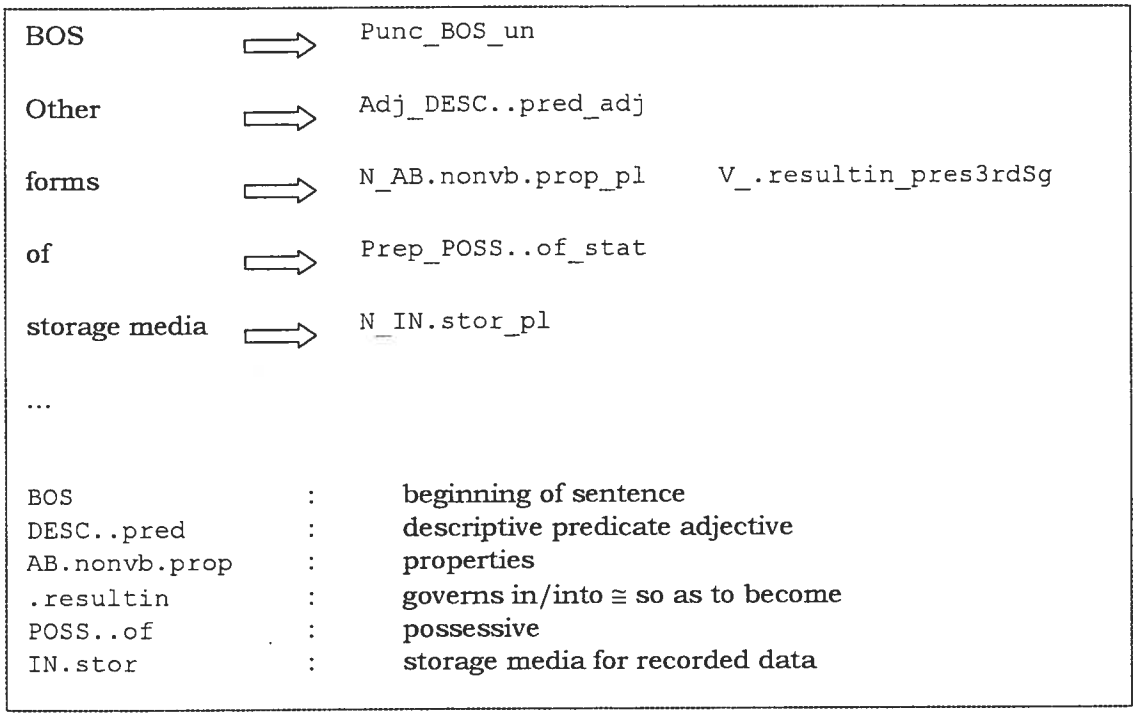


Figure 3-5 Lecture du dictionnaire

La phrase convertie est maintenant envoyée à l'analyseur du système pour être interprétée par les règles de la grammaire. Elles sont plusieurs milliers réparties dans les six modules employés dans l'analyse et le transfert syntaxique d'une phrase donnée. Les règles de l'analyseur décrivent des relations syntaxiques de surface entre les unités lexicales de la langue source. (Celles des quatre derniers modules contiennent également les procédures nécessaires aux opérations de transfert syntaxique.) Ces relations sont exprimées sous la forme de combinaisons

d'unités sémantico-syntaxiques représentées aux différents niveaux de la hiérarchie SAL. Les relations décrites seront ainsi très génériques (*N Prep N*) ou très spécifiques (*free of charge*).

Ainsi qu'il a été mentionné plus haut, les deux premiers modules de l'analyseur (*Resolution modules* ou RES) ont pour tâche principale la désambiguïsation des homographes syntaxiques. Étiqueteur syntaxique du système, RES maintient également une représentation structurelle de la phrase (illustrant les diverses propositions dont elle est composée). La sortie de RES est une chaîne d'unités sémantico-syntaxiques complètement désambiguïsée au niveau syntaxique.

Les quatre modules suivants (*Transfer modules* ou TRAN) sont consacrés à l'analyse de la phrase source et à la construction de l'arbre cible équivalent. Les quatre modules de TRAN collaborent à la construction de l'arbre syntaxique source en partant des feuilles de l'arbre. Lorsqu'un syntagme est identifié à l'un des trois premiers niveaux d'analyse, le dépendant syntaxique est concaténé avec son gouverneur : il ne figure plus dans l'arbre produit par ce niveau. La caractérisation sémantico-syntaxique du gouverneur est modifiée : elle rend compte du nouveau statut de cette unité. Les nœuds restants (non concaténés) à la fin de chaque analyse sont envoyés au niveau suivant.

Alors que les deux premiers modules (TRAN1-2) s'intéressent aux groupes syntaxiques de premier niveau (nominal, verbal, prépositionnel), les deux modules suivants (TRAN3-4) se spécialisent dans les relations intraphrastiques (relations fonctionnelles existant entre les groupes

syntaxiques identifiés auparavant). Les fonctions des quatre modules de TRAN sont *grosso modo* les suivantes[1].

1. TRAN1 identifie et analyse les dépendants linéarisés à gauche du groupe nominal (déterminants, modificateurs adjectivaux, etc.).

Other forms of storage media...

2. TRAN2 identifie les dépendants linéarisés à droite du groupe nominal (compléments du nom, relatives, etc.).

Other forms of storage media...

3. TRAN3 analyse la coordination de groupes nominaux. TRAN3 examine également la structure actancielle de chaque prédicat et détermine l'attachement final du groupe prépositionnel.

... can also be accessed through optional devices

4. TRAN4 finalise l'examen des relations entre les diverses propositions de la phrase et celles existant à l'intérieur de chaque proposition (rôles syntaxiques majeurs, autres

---

[1] Nous présentons seulement les opérations d'analyse.

compléments du verbe, etc.). TRAN4 finalise également la représentation de la structure cible associée à l'arbre syntaxique source.

La sortie des trois premiers modules de TRAN (TRAN1-3) est une représentation intermédiaire de la structure source sous la forme d'une chaîne d'unités sémantico-syntaxiques (les unités associées aux nœuds gouverneurs identifiés par chacun de ces trois premiers modules). Les trois premiers modules de TRAN produisent également une représentation intermédiaire de la structure cible (la structure cible des dépendants syntaxiques concaténés dans chaque module). Les représentations source et cible d'un module sont fournies au module suivant. L'arbre source final est produit à la fin de TRAN3 : tous les nœuds gouverneurs de la phrase ont été identifiés à la fin de TRAN3. La sortie du dernier module (TRAN4) consiste donc en la seule structure cible finale envoyée au générateur morphologique.

C'est sur l'arbre produit à la fin de TRAN3 que nous avons basé l'extraction des combinaisons verbe + nom qui fait l'objet de la présente recherche. L'arbre syntaxique d'une phrase anglaise à la fin de TRAN3 est représentée dans la Figure 3-6.

Magnetic disks provide direct access to data.

```
[
      [Punc BOS   BOS_comma
      ]
      [NP   disk  IN.stor_plwadj
      ]
      [V.   provide     DITR.provide_base
      ]
      [NP   access      OB.undif_sgwadj
      ]
      [Punc EOS   EOS_un
      ]
]
```

Figure 3-6 Arbre syntaxique d'une phrase anglaise

Ainsi que l'illustre la Figure 3-6, l'arbre syntaxique produit par TRAN3 ne représente que les nœuds gouverneurs de cette courte phrase : il s'agit des unités sémantico-syntaxiques associées aux unités lexicales DISK, PROVIDE et ACCESS. Les patrons linguistiques développés pour l'acquisition des collocations de la langue de l'informatique s'appuieront donc sur le formalisme de représentation des unités lexicales du système Logos et plus particulièrement sur la caractérisation morphosyntaxique des unités sémantico-syntaxiques figurant dans les arbres de TRAN3. Nous exploiterons également la caractérisation sémantico-syntaxique des verbes de contrôle, *like*, *want*, *keep*, *stop*, *help*, etc. Nous terminons donc cette présentation du système de TA Logos par la description de SAL, le système de représentation des unités lexicales qui forme la base de ce système.

3.2.2.2 Description de SAL

Base ontologique et structurale du système Logos, SAL est l'ensemble des classes sémantico-syntaxiques qui servent à décrire les unités lexicales de la langue dans le dictionnaire de ce système[2]. Les classes sémantico-syntaxiques de SAL encodent les caractéristiques sémantiques et syntaxiques communes à un groupe d'unités lexicales à l'intérieur de chaque partie du discours. Elles modélisent, pour chaque groupe de lexies désignées, la caractéristique sémantique générale qui permet de mieux rendre compte de leurs propriétés de combinatoire.

Les classes SAL différencient par exemple deux types d'objets parmi les unités nominales qui dénotent des entités : les objets disposant d'une autonomie de fonctionnement (les objets complexes tels que les machines, appareils, organes, etc. désignés par l'étiquette `agentive`) et ceux qui sont nécessairement actionnés (les objets utilitaires, éléments constitutifs, parties du corps, etc. désignés par l'étiquette `functional`). Ces caractéristiques sémantiques différentes (représentées par les étiquettes qui désignent les deux types d'objets) permettent de désambiguïser les exemples suivants, c'est-à-dire de déterminer le rôle syntaxique du groupe prépositionnel, argument dans le premier exemple, circonstant dans le deuxième.

(21) *An X was installed by the computer* (`agentive`)

(22) *An X was installed by the window* (`functional`)

---

[2] Scott (1999) présente les principes sous-jacents à l'élaboration de SAL.

SAL comprend trois niveaux de description sémantico-syntaxique organisés en ordre de spécificité croissante : *superset*, *set* et *subset*. La Figure 3-7 reproduit une partie de la taxonomie SAL des unités nominales anglaises.

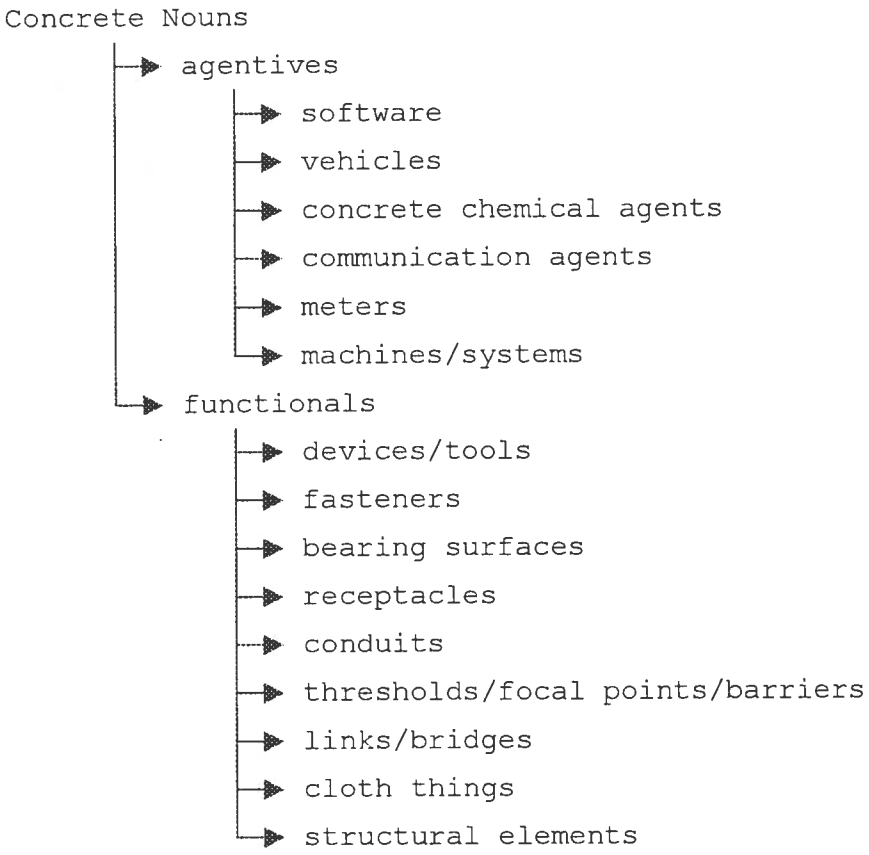


Figure 3-7 Taxonomie SAL des noms anglais

Les unités lexicales enregistrées dans le dictionnaire du système de TA sont donc décrites à chacun des niveaux de la hiérarchie SAL. L'unité terminologique COMPUTER fait ainsi partie des classes `Concrete Nouns`, `agentives` et finalement `machines/systems`. Ces classes servent

également à décrire un très grand nombre de termes apparentés tels que DRIVE, PRINTER, SCANNER, SERVER, TERMINAL, etc.

Ainsi que l'illustre la Figure 3-7, la taxonomie SAL des noms anglais est essentiellement sémantique. Rappelons qu'elle sert principalement à établir les liens de dépendance syntaxique entre ces unités et les autres lexies de la langue (prépositions, verbes, etc.). Celle des verbes et des adjectifs privilégie la description syntaxique (caractérisation de la structure actancielle et du régime du prédicat considéré). Elle identifie ainsi, parmi les verbes ditransitifs, les quatre classes suivantes (correspondant à quatre schémas de réalisation différents) :

1. `dispense-type (requires to)`
   They contributed money to good causes.

2. `give-type (optional to)`
   She gave him the book.
   She gave the book to him.

3. `fetch-type (optional for)`
   He bought her a ring.
   He bought a ring for her.

4. `provide-type (optional to, with)`
   They furnished us the answers.
   They furnished the answers to us.
   They furnished us with the answers.

Les classes ci-dessus représentent le verbe anglais au niveau intermédiaire de la hiérarchie SAL, le niveau le plus spécifique étant ici réservé à l'unité lexicale elle-même. Les verbes font ainsi l'objet d'une codification unique dans le dictionnaire de Logos : ils sont traités de la

même façon que les lexies appartenant aux classes fermées, prépositions, articles, etc.

3.2.2.3 Comparaison de SAL avec l'étiquetage sémantique du DiCo

L'intégration de la description sémantique et syntaxique des unités lexicales réalisée par SAL est spécifique à Logos. Elle différencie le système de TA des autres systèmes de traitement automatique de la langue (TAL) qui séparent habituellement ces deux niveaux de description linguistique. Un modèle de description des unités lexicales qui sépare formellement les deux types d'information est actuellement implémenté dans un dictionnaire informatisé, conçu spécifiquement pour le TAL. Il s'agit du DiCo, la version informatisée du *Dictionnaire explicatif et combinatoire* (DEC), développé depuis plusieurs années par Igor Mel'čuk et Alain Polguère à l'OLST[3]. L'un des aspects les plus intéressants de l'informatisation du DEC réside dans le développement d'un système d'étiquettes afin de modéliser la description sémantique des lexies (Polguère 2000; Polguère 2003c). Le système d'étiquettes élaboré dans le DiCo prend donc en charge la partie centrale de l'article d'une unité lexicale dans le DEC : sa définition lexicale.

L'étiquetage sémantique du DiCo réalise une adaptation de la définition lexicale du DEC en accord avec l'état de l'art en TAL. Il permet une description sémantique simplifiée de l'unité lexicale, encodant la paraphrase minimale de son sens, c'est-à-dire la composante sémantique

---

[3] Une des premières descriptions du DiCo figure dans Mel'čuk *et al.* 1995.

centrale de sa définition lexicale — son genre prochain ou hyperonyme le plus proche. L'unité lexicale ANGOISSE est ainsi décrite à l'aide de l'étiquette `sentiment`, ÉLOQUENCE l'est par l'étiquette `faculté`.

Ces quelques exemples permettent d'illustrer la caractéristique fondamentale de l'étiquetage du DiCo. Les étiquettes sémantiques du DiCo sont des lexies de la langue ou des expressions linguistiques libres (par exemple `son expressif`). Cette caractéristique des étiquettes sémantiques est directement liée à l'approche lexicographique développée dans le DiCo — la lexicologie explicative et combinatoire — approche qui vise la seule description linguistique des unités lexicales et non le rapport qu'elles peuvent entretenir avec un concept sous-jacent.

Les étiquettes sémantiques du DiCo sont toujours des unités nominales. Cette caractéristique reflète également l'approche purement linguistique du DiCo. L'étiquetage sémantique modélisant le sens lexical, il importe en effet de représenter par la même étiquette deux unités lexicales au contenu sémantique identique, même si elles appartiennent à des parties du discours différentes. Par exemple, les unités lexicales CRAINTE et CRAINDRE sont décrites au moyen de la même étiquette (`sentiment`). Chaque étiquette a cependant son équivalent dans les autres parties du discours, obtenu par exemple en combinant l'étiquette avec son verbe support prototypique : *sentiment* → *éprouver un sentiment.*

En plus de décrire le sens des unités lexicales, les étiquettes sémantiques du DiCo remplissent deux autres fonctions : elles classifient les unités lexicales (elles doivent pouvoir décrire au moins deux unités) et structurent hiérarchiquement le lexique de la langue. Les étiquettes

décrivant des hyperonymes immédiats, elles se prêtent naturellement à une structuration hiérarchique dans laquelle chaque étiquette mère représente la paraphrase minimale de ses étiquettes filles.

La dernière caractéristique des étiquettes sémantiques ressort de ce qui vient d'être dit : l'étiquetage sémantique réalisé dans le DiCo est nécessairement lié à une langue donnée. Il se distingue en cela des descriptions de type ontologique, à valeur plus ou moins universelle, recherchées par la majorité des systèmes de traitement automatique et particulièrement par les systèmes de TA.

La principale différence entre le système d'étiquettes du DiCo et celui de Logos réside dans le niveau de description lexicale représenté dans ces deux systèmes. Les étiquettes du DiCo offrent une caractérisation purement sémantique des lexies alors que les classes SAL associent les caractéristiques sémantiques et syntaxiques des unités lexicales qu'elles décrivent. Les classes sémantico-syntaxiques ont été élaborées de façon empirique, inductive pour faciliter avant tout l'analyse syntaxique des phrases. Ainsi qu'il a été vu, différentes propriétés ont été élaborées pour les différentes parties du discours. La caractérisation des verbes et des adjectifs privilégie la description syntaxique (description de la combinatoire syntaxique de ces unités lexicales), celle des noms (et des adverbes) est sémantique[4].

---

[4] Une caractérisation plus fonctionnelle existe pour les noms dérivés de verbes. Elle permet notamment d'encoder le comportement syntaxique de ces unités lexicales.

La taxonomie SAL des noms est plus conceptuelle que sémantique. Elle encode des régularités de comportement découlant d'une propriété sémantique commune à un niveau relativement abstrait (voir par exemple la différence entre les `Mass Nouns` et les `Concrete Nouns`, deux superset de la taxonomie nominale). Les étiquettes SAL ont pour cette raison une valeur strictement dénominative — il ne s'agit pas de lexies comme dans le DiCo. C'est cette caractéristique de la taxonomie des noms qui permet de décrire celle-ci comme une interlangue. Pour les autres parties du discours, SAL, comme le système d'étiquettes du DiCo, est lié à une langue particulière (il existe ainsi deux versions de SAL, pour les deux langues source du système, l'anglais et l'allemand).

L'étiquetage sémantico-syntaxique des unités lexicales de l'anglais implémenté dans le système Logos est essentiellement pratique, visant avant tout à réduire la complexité inhérente à la description systématique et complète de la langue. Moins de mille classes ont été élaborées pour décrire le lexique de l'anglais. La majorité d'entre elles identifient des propriétés syntaxiques — valence et régime des verbes et adjectifs. C'est sur ces classes que nous baserons l'extraction des collocations de la langue de l'informatique. Elles nous serviront à définir les patrons morphosyntaxiques de Colex, l'outil développé dans le cadre de cette recherche.

La description de SAL termine la présentation des outils utilisés pour traiter le corpus d'étude en vue de son exploitation par Colex. Nous passons donc maintenant à la partie centrale de la thèse, la description de la méthode développée pour extraire les collocations de la langue de l'informatique et la présentation des résultats obtenus.

4 Élaboration d'une méthode originale de repérage des collocations

4.1 Introduction

Ce chapitre présente la méthode de repérage des collocations verbe + nom de la langue de spécialité développée dans le cadre de cette recherche. Méthode hybride telle que développée dans les travaux présentés auparavant (Smadja 1993; Grefenstette et Teufel 1995; Lin 1998; Kilgarriff et Tugwell 2001; Goldman *et al.* 2001), elle combine les approches symbolique et statistique à l'acquisition des collocations.

Le programme d'acquisition de collocations que nous avons développé (Colex) repose avant tout sur une analyse linguistique des phénomènes collocationnels, plus particulièrement sur le modèle formel de description de ces phénomènes de la théorie Sens-Texte. Suivant ce modèle, nous considérons la collocation comme une expression linguistique, la réalisation, au niveau syntaxique, d'une relation lexicale particulière. Nous avons donc développé une grammaire d'acquisition de collocations, constituée d'un ensemble de patrons qui identifient les structures syntaxiques caractéristiques des collocations verbales à l'intérieur d'un corpus arboré de 600 000 occurrences. Ces règles sont formulées sur le modèle des règles d'analyse du système de TA qui a fourni les arbres, en utilisant le même langage symbolique. Lorsque les règles de Colex sont appariées aux arbres du corpus, les paires composées du verbe et de l'un de ses dépendants syntaxiques (les paires décrivant une relation actancielle particulière) sont extraites de chaque arbre.

Nous appliquons ensuite un test statistique afin de dégager les collocations de l'ensemble des combinaisons extraites par Colex. Les mesures de la statistique lexicale permettent en effet d'évaluer la force de la relation entre les deux membres d'une combinaison lexicale donnée et de distinguer les collocations (lien lexical fort) des associations libres (lien lexical trivial). Nous avons ainsi évalué les performances de deux mesures sur les combinaisons extraites par Colex. Il s'agit de l'information mutuelle et du coefficient de vraisemblance.

Voici comment nous avons procédé : suivant la théorie Sens-Texte, nous considérons que le nom forme la base des collocations verbe + nom, il sélectionne le verbe pour exprimer un sens et une relation syntaxique particulière, selon la collocation, celle de son premier, deuxième ou troisième actant syntaxique[1]. Les trois types de collocations sont représentées par exemple par les fonctions lexicales supports **Func**, **Oper** et **Labor**. Par rapport aux verbes de notre corpus, les termes nominaux (les bases potentielles des combinaisons extraites) doivent donc remplir l'un des trois rôles syntaxiques suivants :

Sujet : *Number-crunching **programs** can run very slowly.*

Premier complément : *Run the **program** to make sure it works.*

---

[1] Nous considérons ici les cas les plus courants de collocations verbales. Il existent également des cas plus rares où le verbe sélectionne le nom. Ce dernier représente alors un modificateur comme dans l'expression *promettre monts et merveilles.*

Deuxième complément : *Every print command sends data to the* ***program***.

Nous définissons donc trois types de règles spécialisées dans le repérage des collocations de type **Func**, **Oper** ou **Labor**. Nous repérons dans les groupes verbaux du corpus informatique une relation syntaxique particulière à la fois. Nous identifions ainsi des paires d'unités lexicales, ce qui nous permet d'appliquer le test statistique et de mesurer la solidité du lien entre le verbe et le terme dans une position syntaxique donnée (en postulant que ce test permettra de vérifier le caractère collocationnel d'une relation syntaxique donnée).

L'approche mixte au repérage des collocations adoptée ici se justifie de la façon suivante : les premières expériences d'extraction automatique (Berry-Rogghe 1973; Choueka *et al.* 1983; Church et Hanks 1989) ont montré les limites des approches basées uniquement sur les statistiques. Les programmes statistiques, qui basent la définition de la collocation sur la seule fréquence de cooccurrence, identifient différents types de combinaisons (les associations entre *set* et *off* ou entre *doctor* et *nurse* ramenées par l'un des premiers programmes développés), compromettant par là-même l'utilité des résultats obtenus. Dans les applications développées depuis ces premiers essais, le texte servant de support à l'extraction est enrichi de données linguistiques (parties du discours et relations syntaxiques entre les unités lexicales de la phrase). Les collocations sont alors extraites sur la base de critères morphosyntaxiques. Mais les critères morphosyntaxiques ne suffisent pas pour identifier les collocations d'une langue. En effet, les combinaisons retenues sur ces seules bases ne sont pas toutes des collocations. Un

grand nombre d'entre elles sont formées librement, uniquement selon leur sens individuel (considérer par exemple les combinaisons *[to] own a computer*, *[to] exchange a program* ou *[to] need memory*, toutes trois extraites par Colex). C'est ici que les mesures d'association de la statistique lexicale s'avèrent indispensables. En attribuant un score à chaque combinaison, elles permettent un ordonnancement des combinaisons et font ressortir les véritables collocations.

Les fonctions lexicales de la TST (sur lesquelles nous basons le modèle d'extraction de collocations présenté ici) modélisent des relations syntaxiques dites profondes postulées dans le fonctionnement syntaxique de toutes les langues : relations actancielles de base, modification, coordination, apposition. La réalisation syntaxique en surface de ces relations est cependant spécifique à chaque langue (Mel'čuk *et al.* 1995). Le modèle d'extraction de collocations développé dans le cadre de cette recherche (les patrons morphosyntaxiques d'extraction) doit donc être adapté à chaque nouvelle langue traitée. Il faut également disposer d'un analyseur syntaxique de cette langue. Les règles développées pour l'anglais sont cependant suffisamment abstraites pour pouvoir être facilement adaptées à une autre langue. Finalement, bien que développée sur un corpus technique, la méthode présentée ci-dessous est facilement portable d'un corpus à un autre. Toutefois, nous ne pouvons avancer que les résultats seront les mêmes pour un corpus d'une nature très différente : corpus littéraire par exemple.

La première section de ce chapitre présente le programme d'extraction de collocations (Colex), les règles et processus mis en œuvre pour extraire les combinaisons verbe + nom d'un corpus arboré de

600 000 occurrences. Cette section se divise en trois sous-sections : la première sous-section est consacrée à la description de la grammaire (les règles d'extraction des combinaisons verbe + nom); la deuxième sous-section explique le fonctionnement du programme et les processus additionnels mis en place pour optimaliser le calcul des fréquences nécessaires aux tests statistiques; la troisième sous-section évalue, quant à elle, les résultats obtenus à cette étape de l'extraction (avant application des tests statistiques).

La deuxième section de ce chapitre est consacrée aux techniques de filtrage mises en place pour éliminer les combinaisons libres de la liste des combinaisons retenues. Après avoir décrit le programme développé pour calculer le score de chaque combinaison extraite selon l'une ou l'autre des mesures statistiques considérées (information mutuelle et coefficient de vraisemblance), nous présentons les résultats de l'évaluation de ces deux mesures. Cette dernière sous-section termine la présentation de la méthode d'extraction de collocations développé dans le cadre de la présente thèse.

4.2 Le programme Colex d'extraction des paires verbe + nom

Colex, le programme d'extraction des collocations verbe + nom développé dans le cadre de cette recherche, fait deux choses : il extrait les paires verbe + nom (relation verbe + actant syntaxique donnée) et il totalise

la fréquence (le nombre d'occurrences) de chaque paire extraite[2]. Pour extraire les paires, le programme utilise une grammaire, un ensemble de règles, qui

1. définissent des relations syntaxiques de surface ;
2. spécifient, pour chaque structure identifiée, la combinaison verbe + nom devant être extraite (inscrite dans un fichier) et comptée ;
3. rétablissent l'ordre canonique des compléments correspondant à la relation en syntaxe profonde.

4.2.1 Grammaire de Colex

Afin d'extraire les collocations verbales de la langue de l'informatique, nous avons défini un ensemble de règles qui représentent les structures syntaxiques de surface du groupe verbal anglais. Les règles de Colex sont spécialisées dans l'extraction d'une relation actancielle particulière dans chacune des structures de surface décrites.

Nous avons donc trois types de règles pour les trois types de combinaisons recherchées : les combinaisons formées du verbe et de son

---

[2] Le programme présenté dans les pages qui suivent a été développé en collaboration avec Mike Dillinger, responsable du développement linguistique à Logos à l'époque où nous avons entrepris ce travail. Nous n'aurions pas pu mener celui-ci à terme sans l'aide et le soutien constant de Mike. Nous profitons de cette occasion pour le remercier encore une fois.

premier, deuxième et troisième actant respectivement, ou, en termes de relations syntaxiques de surface, les combinaisons formées du verbe et de son sujet, premier complément et deuxième complément respectivement. La relation verbe + troisième complément, moins fréquente, n'a pas été considérée ici.

Ces règles sont stockées dans trois fichiers différents, trois grammaires utilisées à tour de rôle par le programme d'extraction pour l'acquisition des principales relations actancielles.

Dans la sous-section suivante, nous examinons la microstructure d'une grammaire de Colex. Nous décrivons en détail la partie conditionnelle des règles d'acquisition et le métalangage développé pour représenter les structures syntaxiques de surface de l'anglais. La deuxième sous-section s'intéresse, quant à elle, à la macrostructure d'une grammaire et présente brièvement les différentes structures syntaxiques qui y sont décrites[3].

4.2.1.1 Description d'une règle de Colex

Les règles d'acquisition des combinaisons verbe + nom sont composées de deux parties : une partie conditionnelle qui décrit une structure syntaxique de surface particulière et une partie action qui spécifie les éléments à extraire dans l'arbre correspondant. Cette deuxième partie calcule également la fréquence (le nombre d'occurrences) de chaque

---

[3] La liste complète des règles d'extraction des combinaisons verbe + premier complément est donnée dans l'Annexe B.

combinaison extraite. Dans la description qui suit, nous nous intéresserons avant tout à la partie conditionnelle d'une règle.

Les structures syntaxiques de surface représentées dans la partie conditionnelle d'une règle sont définies sur la base des arbres produits par l'analyseur de Logos. Plus particulièrement, elles sont formatées sur la base des représentations fournies par l'analyseur à la fin de TRAN3, le module qui finalise l'analyse d'une phrase donnée (cf. Figure 3-3, p. 134, du chapitre précédent). Nous reproduisons ci-dessous l'arbre syntaxique associé à une phrase anglaise.

The next field contains the complete name of the Newsgroup.

```
[
      [Punc BOS   BOS_comma
      ]
      [NP   field PL.undif_sgwdefDet
      ]
      [V    contain    contain_pres3rdSg
      ]
      [NP   name  ap_sgwdefDet
      ]
      [Punc EOS   EOS_un
      ]
]
```

Figure 4-1 Arbre syntaxique produit par l'analyseur de Logos à la fin de TRAN3

Ainsi que l'illustre la Figure 4-1, l'arbre syntaxique produit par l'analyseur de Logos à la fin de TRAN3 se présente comme la liste des unités sémantico-syntaxiques associées aux nœuds gouverneurs de la phrase de départ ; les unités marquées BOS et EOS sont insérées

automatiquement pour signaler le début et la fin de chaque phrase analysée. Rappelons qu'une unité sémantico-syntaxique est une représentation interne qui encode, pour chaque unité lexicale stockée dans le dictionnaire du système, trois types d'information : partie du discours, caractérisation sémantico-syntaxique et morphosyntaxique. Ces informations sont modifiées au cours de l'analyse afin de

1. contrôler l'appariement des unités sémantico-syntaxiques avec les règles de l'analyseur et

2. construire l'arbre syntaxique correspondant à chaque niveau d'analyse (TRAN1-3).

En construisant l'arbre syntaxique représenté à la Figure 4-1, l'analyseur normalise les unités sémantico-syntaxiques de départ : à la fin de TRAN3, il s'agit uniquement de noms, de verbes, d'auxiliaires et d'éléments frontières (signes de ponctuation, conjonctions, etc.). La classe nominale fait ainsi office de « classe fourre-tout ». À la fin de TRAN3, en plus des noms et groupes nominaux, elle regroupe également les adverbes et adjectifs dépendants du verbe et tous les groupes prépositionnels (actants ou circonstants). La normalisation de ces unités lexicales ou groupes d'unités lexicales sous la classe nominale permet, entre autres, de généraliser les transformations dans le dernier module de l'analyseur (TRAN4).

La partie conditionnelle d'une règle de Colex est donc basée sur cet arbre syntaxique normalisé. On peut se la représenter comme une matrice de valeurs à unifier à l'arbre d'entrée. En abscisses, on trouve une suite ordonnée d'éléments à apparier aux nœuds gouverneurs de l'arbre

d'entrée, en ordonnées, la liste des valeurs définies pour chaque élément de la suite (sur la base des valeurs des unités sémantico-syntaxiques de Logos). Nous donnons ci-dessous l'exemple de la première règle de la grammaire d'acquisition des combinaisons sujet + verbe (R1).

```
(R1) if(BOP (adv) anyNoun (anyAux) vt_active (adv) (n_obj) (adv) EOP
           then {anyNoun + vt_active;}
     endif
```

The next field$_{[anyNoun]}$ *contains*$_{[vt\_active]}$ *the complete name of the Newsgroup.* ⟹ `field + contain(vt)`

Ainsi que l'illustre l'exemple ci-dessus, les règles utilisent deux moyens d'identification des différentes relations actancielles à l'intérieur d'une structure syntaxique donnée : l'ordre d'apparition des éléments dans la règle (qui correspond à l'ordre des mots de la phrase anglaise) et l'étiquetage morphosyntaxique (qui correspond à l'étiquetage des unités sémantico-syntaxiques de l'arbre). Les parenthèses permettent de baliser les éléments optionnels pour l'appariement d'une règle avec un arbre donné.

Comme les unités sémantico-syntaxiques de Logos, les éléments (ou variables) des règles de Colex permettent la caractérisation des unités lexicales aux trois niveaux suivants :

1. partie du discours,

2. caractérisation sémantico-syntaxique,

3. caractérisation morphosyntaxique.

Par exemple, les variables `anyNoun` et `anyAux` de la règle (R1) ci-dessus se réécrivent ainsi :

```
AnyNoun=N _ any _ any
AnyAux=Aux _ any _ any
```
[4]

Elles s'unifient par exemple avec les expressions suivantes :

(23) *The installer*$_{[anyNoun]}$ *does*$_{[anyAux]}$ *check...*

(24) *You*$_{[anyNoun]}$ *can*$_{[anyAux]}$ *save...*

(25) *Any other C compiler*$_{[anyNoun]}$ *will*$_{[anyAux]}$ *probably come with...*

Les valeurs (`N` ou `Aux`) apparaissant à l'intérieur des deux variables ci-dessus correspondent aux valeurs définies dans Logos pour l'unité sémantico-syntaxique équivalente. La caractérisation à chacun des trois niveaux d'une variable peut également correspondre à une liste de valeurs, liste qui sera elle-même généralement représentée par une nouvelle variable. C'est le cas des deux listes suivantes définies pour caractériser

---

[4] Le trait de soulignement assure la concaténation des valeurs assignées aux variables.

un verbe aux niveaux sémantico-syntaxique et morphosyntaxique respectivement :

```
trans=monoTR|preclausal|complexTR ...
active=base|am|pres3rdSg ...
```

La variable correspondant au verbe se réécrivant alors

```
vt_active=V _ trans _ active,
```

elle s'unifiera avec les verbes transitifs des deux premiers exemples ci-dessus — *check*$_{[vt\_active]}$ et *save*$_{[vt\_active]}$ — et non avec le verbe du troisième exemple *come*$_{[vi\_active]}$.

L'usage de la variable `trans` (et de la variable `intr`) pour décrire le verbe recherché à l'intérieur de chaque règle permet au programme d'extraire et de compter séparément les acceptions transitives et intransitives des verbes du corpus. Ainsi, le programme extrait les combinaisons sujet + verbe des exemples suivants comme deux combinaisons différentes :

This program [anyNoun] *does not execute* [vt_active] *any FPU instruction.* ⟹ `program + execute(vt)`

Several programs [anyNoun] *can execute* [vi_active] *concurrently.* ⟹ `program + execute(vi)`

Le verbe sera également caractérisé au niveau morphosyntaxique. Des variables existent ainsi pour définir un verbe au passif (`vt_passive`), à l'impératif (`vt_imperative`), à l'infinitif (`vt_infinitive`), à l'infinitif passif (`vt_inf_pass`), etc.

Parmi les variables définies pour les unités nominales (plus particulièrement les actants syntaxiques d'un verbe), `anyNoun` sert à représenter le sujet grammatical. C'est la seule variable, parmi celles servant à désigner les actants syntaxiques du verbe, à ne pas faire l'objet d'une caractérisation plus précise. Le rôle grammatical de l'unité lexicale correspondante est ici entièrement déterminé par la position de la variable à l'intérieur de la règle. Nous avons également `n_obj` pour tout (premier) complément et `n_iobj` pour tout complément autre que le premier. Ces deux variables sont représentées dans la règle ci-dessous (R2) qui extrait une combinaison verbe + premier complément (`n_obj`).

```
(R2) if(BOP (adv) anyNoun (anyAux) vt_active (adv) n_obj (n_iobj) (adv) EOP)
            then {vt_active + n_obj;}
     endif
```

The user activates[vt_active] *the program*[n_obj]. ⟹ `activate(vt) + program`

Les variables `n_obj` et `n_iobj` sont différenciées uniquement au niveau de la caractérisation morphosyntaxique, seule caractérisation permettant de distinguer les groupes nominaux des groupes prépositionnels à la fin de TRAN3. (Rappelons que la partie du discours d'un groupe prépositionnel à la fin de TRAN3 est celle du nom.) La caractérisation morphosyntaxique de `n_obj` inclut donc les valeurs possibles des groupes nominaux et prépositionnels alors que celle de `n_iobj` ne contient que les valeurs des groupes prépositionnels.

```
(R3) if(… anyNoun (anyAux) vt_active (adv) n_obj n_iobj (n_iobj) …)
              then {vt_active + n_iobj;}
     endif
```

This does not mean one should download [vt_active] *software from just any website* [n_iobj]. ⟹ `download(vt) + from + website`

Comme nous l'avons déjà signalé, l'utilisation de parenthèses permet d'indiquer le statut obligatoire ou facultatif d'une variable pour l'appariement d'une règle avec un sous-arbre donné. Ainsi, dans la règle (R2) reproduite ci-dessous, le deuxième complément (`n_iobj`) est marqué facultatif. La règle sera exécutée (et la combinaison verbe + premier complément sera extraite) que le verbe ait ou non un deuxième complément (que celui-ci soit ou non réalisé dans la phrase).

```
(R_) if(BOP (adv) anyNoun (anyAux) vt_active (adv) n_obj (n_iobj) (adv) EOP)
            then (vt_active + n_obj;)
    endif
```

A 2x card transfers$_{[vt\_active]}$ *data*$_{[n\_obj]}$ *twice per clock cycle.* ⟹ `transfer(vt) + data`

A personal computer may transfer$_{[vt\_active]}$ *data*$_{[n\_obj]}$ *from disk to CPU, from CPU to memory, or from memory to the display adapter.* ⟹ `transfer(vt) + data`

La variable facultative pour un appariement donné (le deuxième complément dans les exemples ci-dessus) peut également être répétée : elle s'unifie avec zéro ou plusieurs nœuds gouverneurs comme dans le deuxième exemple. Le marquage de certains éléments comme étant facultatifs permet de généraliser l'appariement des règles (d'apparier une règle donnée au plus grand nombre de phrases possible). Il permet également de limiter la taille des grammaires développées (d'écrire moins de règles). Les autres éléments facultatifs comprennent les adverbes et les auxiliaires.

Les règles sont sensiblement les mêmes d'une grammaire à l'autre ; elles décrivent les mêmes structures syntaxiques de surface avec le même nombre d'éléments. Ce qui change d'une grammaire à l'autre, c'est la

relation actancielle (parmi celles qui sont représentées dans la partie conditionnelle) qui est extraite. L'extraction d'une relation actancielle donnée est encodée dans la partie action de la règle. Cette partie est composée de deux fonctions. La première fonction spécifie quelles unités lexicales (parmi celles ayant été appariées) doivent être extraites de l'arbre et imprimées dans le fichier résultat. La deuxième fonction permet de compter le nombre de fois qu'une combinaison particulière a été extraite. La fréquence de chaque paire extraite est une des données nécessaires au test statistique qui suit.

Après avoir examiné en détail la structure des règles de Colex, nous abordons maintenant la couverture des grammaires de Colex.

4.2.1.2 Description des structures syntaxiques couvertes par une grammaire

Les grammaires de Colex acquièrent les combinaisons verbe + nom dans des représentations de la structure syntaxique de surface des phrases du corpus. Elles doivent donc prendre en charge des structures nombreuses et variées. Chaque grammaire contient ainsi plusieurs dizaines de règles : la grammaire d'acquisition des combinaisons verbe + premier complément compte ainsi 112 règles ; celle spécialisée dans l'acquisition des combinaisons sujet + verbe en compte 34 et celle développée pour acquérir les combinaisons verbe + deuxième complément, 62[5]. Dans la description des structures syntaxiques couvertes par les

---

[5] Rappelons que la liste complète des règles de la première grammaire d'acquisition figure dans l'Annexe B.

grammaires de Colex qui suit, nous ne donnerons donc qu'un seul exemple parmi les règles conçues pour chacune des structures traitées.

Le nombre relativement élevé de règles à l'intérieur des trois grammaires est attribuable en partie au besoin d'assigner différentes caractérisations morphosyntaxiques au verbe (verbe actif, passif, à l'impératif, etc.). Nous avons ainsi défini des règles pour des structures passives qui rétablissent, pour l'extraction, l'ordre actif des compléments. Nous donnons ci-dessous un exemple de règle pour une structure passive.

```
(R4) if(BOP (adv) anyNoun (anyAux) vt_passive (adv) (n_obj) (adv) EOP)
           then {vt_passive + anyNoun;}
     endif
```

The program[anyNoun] *is automatically executed*[vt_passive] *on the local computer.* ⟹ `execute(vt) + program`

Nous avons également écrit des règles pour des verbes à l'impératif et des verbes dans des propositions non tensées : verbes à l'infinitif et au participe présent (gérondifs). Des exemples de ces constructions sont données ci-dessous.

```
(R5) if(BOP vt_imperative (adv) n_obj (n_iobj) (adv) EOP)
          then {vt_imperative + n_obj;}
     endif
```

Type [vt_imperative] *this program* [n_obj] *into a file.* ⟹ `type(vt) + program`

```
(R6) if(BOP vt_infinitive (adv) n_obj (n_iobj) (adv) EOP)
          then {vt_infinitive + n_obj;}
     endif
```

To execute [vt_infinitive] *the queued commands* [n_obj]*, just click the green Go button.* ⟹ `execute(vt) + command`

```
(R7) if(BOP vt_ing (adv) n_obj (n_iobj) (adv) EOP)
          then {vt_ing + n_obj;}
     endif
```

Those systems will wait for any operation to complete before running[vt_ing] *any other program*[n_obj]. ⟹ `run(vt) + program`

Dans les exemples de règles qui précèdent, les propositions infinitives ou gérondives sont de nature circonstancielle ; seules les combinaisons verbe + complément peuvent être extraites. Lorsque l'infinitive (ou la gérondive) est complément du verbe et qu'elle représente un des actants du verbe de contrôle, il est possible d'acquérir également une combinaison sujet + verbe :

```
(R8) if(… anyNoun (anyAux) v_preverbal_act vt_infinitive (adv) n_obj …)
          then {anyNoun + vt_infinitive;}
     endif
```

The program[anyNoun] *wants to read*[vt_infinitive] *something from a disk.* ⟹ `program + read(vt)`

```
(R9) if(… anyNoun (anyAux) v_prevDITR_act (n_dobj) vt_infinitive (adv) n_obj …)
        then {n_dobj + vt_infinitive;}
    endif
```

We told the computer$_{[n\_dobj]}$ *to draw*$_{[vt\_infinitive]}$ *one line.* ⇒ `computer + draw(vt)`

Finalement, nous avons défini un nombre important de règles pour décrire les structures avec verbes coordonnés. Nous donnons en exemple deux de ces règles pour l'extraction d'une combinaison sujet + verbe et verbe + premier complément respectivement : les règles R10 et R11 ci-dessous. Dans l'un et l'autre cas, le verbe extrait est celui qui se trouve le plus éloigné de l'actant syntaxique, sujet ou complément, que la règle a pour but d'acquérir.

```
(R10) if(… anyNoun (anyAux) vt_act conj (anyAux) v_act (adv) n_obj (n_iobj) …)
        then {vt_act + n_obj;}
    endif
```

This is how the microprocessor loads$_{[vt\_act]}$ *and executes the entire operating system*$_{[n\_obj]}$. ⇒ `load(vt) + operating system`

```
(R11) if(…anyNoun (anyAux) v_act (adv) (n_obj) conj (anyAux) vi_act (adv) (n_obj)…)
            then {anyNoun + vi_act;}
      endif
```

The computer$_{[anyNoun]}$ *will click and whir*$_{[vi\_act]}$ *for a few moments.* ⟹ `computer + whir(vi)`

Pour terminer cette présentation des règles de Colex, soulignons le principe directeur qui a guidé le développement de chaque grammaire d'extraction de collocations : extraire le plus grand nombre de combinaisons verbales, même à partir de structures incomplètes.

4.2.2 Fonctionnement de Colex

Ainsi que nous venons de le voir dans la section ci-dessus, Colex dispose de trois grammaires pour extraire les trois types de relations syntaxiques que décrivent les collocations verbe + nom : relation sujet + verbe (grammaire **Subj**), relation verbe + premier complément (grammaire **Obj1**) et relation verbe + deuxième complément (grammaire **Obj2**). Pour extraire une relation actancielle donnée, il suffit donc de préciser la sous-grammaire à utiliser (**Subj**, **Obj1** ou **Obj2**). Pour extraire toutes les combinaisons du corpus, il faut exécuter le programme trois fois et changer la grammaire à chaque exécution.

À chaque exécution du programme, les combinaisons illustrant une relation syntaxique donnée sont réparties dans deux fichiers différents : les acceptions intransitives des vocables verbaux sont imprimées dans un fichier appelé freqsVIOUT et les acceptions transitives vont dans un fichier

appelé freqsVTOUT. Chaque exécution de Colex génère donc deux fichiers. Il s'agit de fichiers au format texte. Les résultats y sont présentés sous forme de liste. Chaque ligne décrit une combinaison lexicale particulière (une paire d'unités lexicales) au moyen des trois informations suivantes : la fréquence de la paire (nombre d'occurrences relevées dans le corpus), le verbe et le nom. L'ordre de présentation des unités lexicales est inversé pour les combinaisons sujet + verbe. L'addition des lignes contenues dans les deux fichiers freqs nous donne le total des paires extraites pour chaque relation :

Obj1 : 19 616 paires ;

Subj : 12 685 paires ;

Obj2 : 7 899 paires.

La Figure 4-2 montre les 20 premières paires (en termes de fréquence) extraites pour la relation verbe + premier complément, relation pour laquelle est extraite le plus grand nombre de paires.

```
f:    96    do(vt)          this
f:    93    see(vt)         figure
f:    70    use(vt)         it
f:    61    store(vt)       data
f:    59    use(vt)         command
f:    49    send(vt)        data
f:    37    do(vt)          thing
f:    36    do(vt)          it
f:    35    take(vt)        look
f:    34    use(vt)         computer
f:    34    run(vt)         program
f:    34    press(vt)       key
f:    32    press(vt)       button
f:    30    transfer(vt)    data
f:    30    use(vt)         program
f:    27    send(vt)        email
f:    27    execute(vt)     instruction
f:    27    take(vt)        advantage
f:    26    have(vt)        problem
f:    26    do(vt)          job
```

Figure 4-2 Extrait du fichier freqsVTOUT des paires verbe + premier complément

Les deux figures suivantes listent les vingt premières paires extraites pour les relations sujet + verbe et verbe + deuxième complément respectivement.

```
f:    301   you      have(vt)
f:    269   you      use(vt)
f:    150   you      see(vt)
f:    116   you      get(vt)
f:    105   you      do(vt)
f:    100   you      find(vt)
f:    99    you      need(vt)
f:    84    it       have(vt)
f:    74    you      know(vt)
f:    70    we       use(vt)
f:    68    you      install(vt)
f:    68    you      create(vt)
f:    62    this     mean(vt)
f:    57    you      learn(vt)
f:    56    you      change(vt)
f:    51    I        have(vt)
f:    50    you      make(vt)
f:    41    system   have(vt)
f:    41    you      want(vt)
f:    40    it       take(vt)
```

Figure 4-3 Extrait du fichier freqsVTOUT des paires sujet + verbe

```
f:    16   load(vt) in          memory
f:    13   cover(vt) in         chapter
f:    10   discuss(vt) in       detail
f:    10   send(vt) to          program
f:    9    send(vt) to          server
f:    9    connect(vt) to       motherboard
f:    8    scroll down(vt) in   order
f:    8    read(vt) in          order
f:    8    base(vt) on          technology
f:    8    store(vt) on         disk
f:    7    connect(vt) to       internet
f:    7    send(vt) to          someone
f:    7    provide(vt) for      user
f:    7    keep(vt) in          mind
f:    6    store(vt) in         memory
f:    6    connect(vt) to       each other
f:    6    store(vt) in         ram
f:    6    convert(vt) in       tone
f:    6    discuss(vt) in       chapter
f:    6    download(vt) from    internet
```

Figure 4-4 Extrait du fichier freqsVTOUT des paires verbe + deuxième complément

Afin d'optimaliser le relevé des fréquences des combinaisons verbe + nom par les règles des grammaires de Colex, nous avons développé

trois sous-programmes qui examinent et éventuellement modifient les mots-formes associés aux nœuds de chaque arbre syntaxique avant que ceux-ci ne soient extraits de l'arbre. Les programmes en question visent à compenser certains effets négatifs de l'analyse syntaxique des phrases du corpus par Logos — dans les deux derniers cas, il s'agit même de « défaire » l'analyse effectuée par le système de TA. Nous décrivons brièvement ces programmes dans les sous-sections suivantes. L'ordre de présentation est celui de l'exécution des sous-programmes à l'intérieur de Colex. Il s'agit des sous-programmes de lemmatisation (verbes), de traitement des verbes à particules et de traitement des groupes prépositionnels.

4.2.2.1 Lemmatisation

Nous avons dû tout d'abord lemmatiser les formes verbales (et certaines formes nominales) qui figuraient dans les arbres d'entrée. En effet, l'analyseur de l'anglais de Logos ne lemmatise pas la majorité des formes fléchies rencontrées dans les textes — seules les formes du pluriel des noms communs et de la troisième personne du singulier des verbes font l'objet d'une analyse morphologique. Les formes fléchies des unités lexicales de l'anglais sont donc enregistrées dans le dictionnaire du système. Elles constituent l'entrée du dictionnaire de chaque unité lexicale décrite dans ce système.

Le dictionnaire stocke jusqu'à trois formes par entrée. Pour les verbes, il s'agit le plus souvent des formes de l'infinitif (forme canonique), du participe présent (forme en *-ing*) et du participe passé (forme en *-ed*). Selon le contexte, c'est l'une ou l'autre de ces formes qui est récupérée lors

de la lecture du dictionnaire et manipulée tout au long de l'analyse. Elle figure donc dans l'arbre passé à Colex.

La représentation des formes fléchies dans les arbres produits par l'analyseur de Logos affectait négativement le calcul de la fréquence d'une paire verbe + nom donnée : chaque forme, infinitif, participe présent ou participe passé d'un verbe étant comptée comme une occurrence séparée. Par exemple, les trois formes de *load* des exemples suivants étaient extraites comme trois combinaisons différentes au lieu d'être analysées comme des occurrences de la seule collocation `load(vt) + into + memory`.

(26) *If you want to continue, you have to load the document into memory again.*

(27) *The computer cannot do anything without first loading an operating system into memory.*

(28) *After the BIOS boot program has loaded the boot record into memory...*

Nous avons donc développé un sous-programme pour remplacer chaque forme verbale fléchie dans l'arbre par la forme de départ (le lemme) du verbe correspondant. Nous remplaçons une chaîne de caractères par une autre. Cette opération (lemmatisation) repose sur une analyse morphologique des formes verbales anglaises. Le remplacement peut porter sur la forme toute entière — le programme consulte en premier une liste de formes irrégulières — ou sur une partie seulement, c'est-à-dire sur le morphème *-ing* ou *-ed* final et la ou les dernières lettres du radical.

4.2.2.2 Traitement des verbes à particules

Le deuxième sous-programme développé pour traiter les mots-formes figurant dans les arbres avant l'extraction des combinaisons verbe + nom a pour objet le traitement des verbes à particules (*phrasal verbs*) : *[to] back up*, *[to] turn on*, *[to] turn off*, etc. Ces expressions sont analysées dans le deuxième module de l'analyseur (TRAN2) au moyen de règles spéciales stockées dans la base de données collocationnelles du système (SemTab).

Lors de l'exécution d'une règle de SemTab pour un verbe à particule, la deuxième unité lexicale (le plus souvent un adverbe) est concaténée avec le verbe — le nœud correspondant est effacé de l'arbre — et la caractérisation SAL de ce dernier est modifiée : le verbe reçoit les valeurs sémantico-syntaxiques qui caractérisent la structure actancielle et le régime du nouveau prédicat et permettent de distinguer celui-ci de la forme trouvée dans le dictionnaire en début d'analyse. Rappelons que, pour un verbe, la valeur assignée au niveau le plus spécifique de la hiérarchie SAL (*subset*) sert à identifier ce verbe de façon unique. C'est cette dernière valeur qui permet de récupérer l'adverbe concaténé et de l'imprimer dans le fichier résultat de Colex. Le sous-programme vérifie le *subset* de chaque forme verbale figurant dans l'arbre. Si la valeur du *subset* est celle d'un verbe à particule, le programme remplace la forme (simple) du verbe dans l'arbre par la séquence verbe + adverbe qui convient.

4.2.2.3 Traitement des groupes prépositionnels (premier et deuxième compléments seulement)

Le dernier sous-programme exécuté avant l'extraction des combinaisons verbe + nom concerne les groupes prépositionnels. Comme il a été mentionné plus haut, les groupes prépositionnels apparaissent comme des noms dans les arbres produits par le dernier module d'analyse (TRAN3). Les groupes prépositionnels qui dépendent du verbe sont analysés dans le dernier module d'analyse — l'une des principales fonctions de TRAN3 est de déterminer le rôle grammatical de ces compléments. Quel que soit son rôle, le groupe prépositionnel a pour nœud gouverneur l'unité sémantico-syntaxique correspondant au nom (objet de la préposition). Les règles d'analyse opèrent la concaténation de la préposition — comme l'adverbe des verbes à particules, elle est effacée de l'arbre. Elles mettent à jour la caractérisation morphosyntaxique du nom (pour indiquer le caractère particulier du groupe) ainsi que sa caractérisation sémantico-syntaxique. On donne ainsi au nom la valeur de *subset* de la préposition. Comme pour les verbes, le *subset* d'une préposition identifie cette préposition de façon unique. L'analyse du groupe prépositionnel est illustrée dans la Figure 4-5.

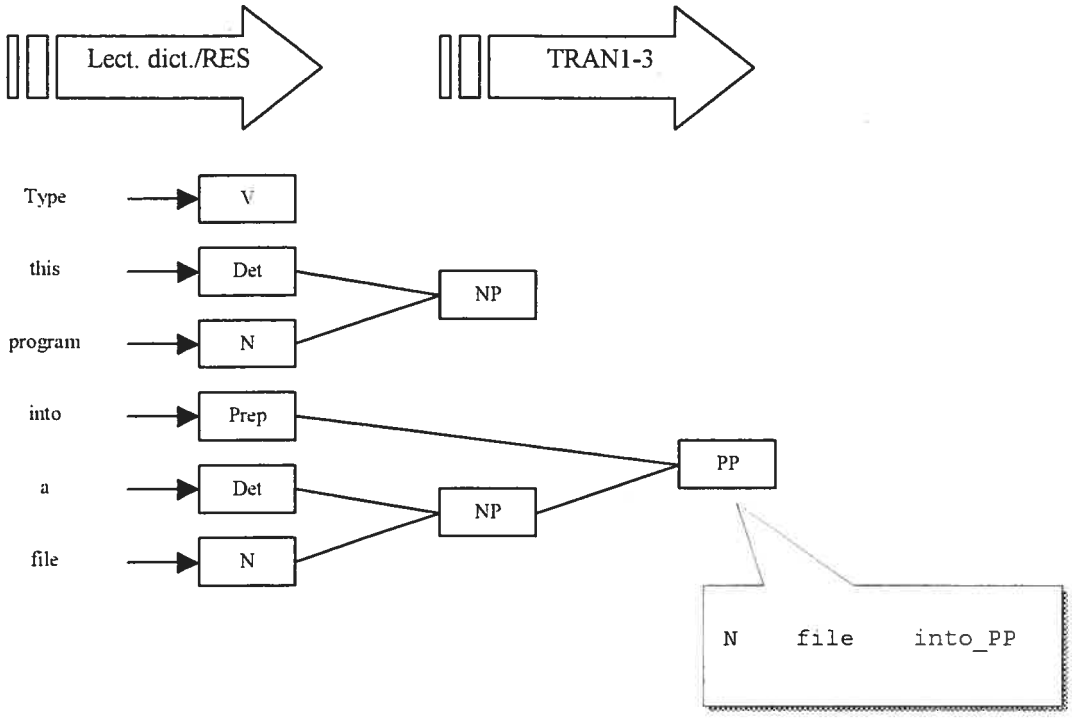


Figure 4-5 Analyse du groupe prépositionnel

La méthode développée pour le traitement des verbes à particules est également appliquée ici : le programme vérifie le *subset* de chaque complément (actant ou circonstant) dans l'arbre. Si celui-ci correspond à l'identificateur d'une préposition anglaise, le programme insère une préposition dans l'arbre. Le traitement des groupes prépositionnels diffère de celui réservé aux verbes à particules de deux façons. Premièrement, le programme n'insère pas la préposition concaténée mais un hyperonyme de celle-ci. Cet hyperonyme est choisi en fonction de la valeur sémantico-syntaxique de la préposition au deuxième niveau de la taxonomie SAL (le niveau du *set* qui réalise un regroupement sémantique des prépositions). Des exemples de valeurs SAL des prépositions anglaises au niveau du *set* sont données ci-dessous.

1. in-type
   amid, amidst, among, amongst, between, in, in between, inside, into, throughout, within

2. on-type
   aboard, on, on top of, onto, upon

3. at-type
   against, along, alongside, alongside of, at, beside, near

4. from-type
   from, off, off of, out of

Deuxièmement, l'hyperonyme (la chaîne de caractères associée) d'une préposition concaténée est insérée auprès du verbe (et non du nom) dans l'arbre syntaxique. Ceci permet de sous-catégoriser, dans les fichiers résultats, les acceptions transitives ou intransitives des verbes selon les prépositions qu'ils régissent et d'optimaliser le relevé des fréquences propres aux verbes pour le traitement statistique à venir. Parmi les acceptions intransitives de *[to] talk*, le programme peut ainsi distinguer et compter séparément les occurrences de *[to] talk* + *about* et de *[to] talk* + *to*. Nous distinguons pareillement, parmi les acceptions transitives de *[to] load*, les combinaisons verbe + deuxième complément *[to] load* + *from* et *[to] load* + *in*.

4.2.3 Évaluation des résultats obtenus

Nous abordons maintenant l'évaluation de l'extraction linguistique des combinaisons verbe + nom de la langue de l'informatique. Pour procéder à cette évaluation, nous avons sélectionné dix unités terminologiques, parmi les plus fréquentes de notre corpus, et examiné les combinaisons lexicales extraites par les trois grammaires de Colex pour chacune de ces unités (un total de 4 941 occurrences). Les dix unités terminologiques (l'ensemble des combinaisons les contenant) serviront également à l'évaluation des différents filtres statistiques dans la dernière section de ce chapitre. Le tableau ci-dessous donne le total des combinaisons extraites par chacune des grammaires de Colex pour les dix termes retenus (les résultats sont exprimés en termes d'occurrences).

	Subj	Obj1	Obj2	Total
COMPUTER	295	323	159	777
DATA	55	557	23	635
DISK	46	117	64	227
DRIVE	61	151	55	267
FILE	111	589	82	782
INFORMATION	31	296	36	363
MEMORY	47	118	76	241
PROGRAM	326	443	105	874
SERVER	198	147	98	443
SOFTWARE	120	174	38	332
Total	1 290	2 915	736	4 941

Tableau 4-I Combinaisons extraites pour les dix termes retenus dans le corpus informatique

Comme le montre le tableau ci-dessus, les termes retenus pour l'évaluation de Colex désignent des concepts fondamentaux du domaine de l'informatique. Il s'agit de vocables monosémiques (à l'intérieur du domaine de spécialité considéré) qui se prêtent à un nombre élevé de combinaisons. Ces deux caractéristiques ont été déterminantes pour retenir un terme pour l'évaluation du programme d'acquisition de collocations. Un autre critère de sélection a consisté à retenir des termes qui désignaient des objets différents du domaine étudié afin d'acquérir une variété de collocatifs verbaux.

Par ailleurs, chaque terme retenu sert de base à de nombreux phrasèmes spécialisés dont il est également l'hyperonyme le plus proche. Donnons pour exemple l'unité lexicale DISK, hyperonyme des deux phrasèmes spécialisés HARD DISK et FLOPPY DISK. Selon les contextes, on la trouvera combinée avec les collocatifs verbaux de l'un ou de l'autre de ses hyponymes (*[to] eject* ou *[to] insert* lorsqu'elle est utilisée à la place de

FLOPPY DISK, *install* ou *mount* lorsqu'il s'agit de HARD DISK). Pour une description adéquate de la cooccurrence verbale de DISK, il serait nécessaire de distinguer ces différents collocatifs des emplois prototypiques de l'unité lexicale.

Pour évaluer les combinaisons lexicales extraites par le programme développé dans le cadre de notre recherche, nous avons utilisé les deux mesures habituelles de rappel et de précision. Dans notre application, la précision mesure la proportion de combinaisons correctes par rapport au nombre de combinaisons retenues alors que le rappel mesure la proportion de combinaisons retenues par rapport à l'ensemble des combinaisons du corpus.

4.2.3.1 Mesure du rappel

Pour mesurer le rappel de l'extraction linguistique des collocations de la langue de l'informatique, nous avons comparé le nombre de concordances obtenues pour un terme donné au nombre de combinaisons extraites pour ce terme. Cette analyse nous a également permis d'évaluer la couverture des différentes grammaires de Colex. Nous avons ainsi sélectionné 178 phrases au hasard parmi les concordances de PROGRAM. Sur les 213 occurrences de combinaisons verbe + nom présentes dans l'échantillon, seules 131 occurrences ont été extraites automatiquement. Le programme a donc ignoré 82 combinaisons, ce qui représente un rappel de 61,50 %. Trois facteurs contribuent à ce faible taux de rappel. Ils sont résumés ci-dessous.

Un peu moins de la moitié des omissions (45,12 %) concernent des combinaisons à l'intérieur de propositions relatives ou d'autres constructions absolues; par exemple :

(29) *This is not a real problem for a program that you use frequently.*

(30) *There is very little that's special about CGI programs executed from within an SSI file.*

Les propositions relatives ne sont pas décrites dans les grammaires de Colex. En effet, aucune indication ne nous permettait de rétablir le lien entre le verbe et l'actant extraposé (l'antécédent du pronom relatif) une fois la proposition analysée par Logos. Nous avons donc été obligée d'exclure les propositions relatives de la couverture syntaxique de Colex.

Le quart environ des omissions (25,61 %) est due à des erreurs d'analyse par Logos. Parmi ces erreurs, celles d'étiquetage des parties du discours et d'analyse du groupe nominal contribuent au plus grand nombre d'omissions. Des exemples de ces deux types d'erreurs sont donnés ci-dessous.

(31) *Environment variables are certainly integral to making your CGI program work.* (erreur sur *work* étiqueté comme un nom)

(32) *Even OS/2's own house maintenance programs can be viewed.* (groupe nominal sujet analysé comme trois unités sémantico-syntaxiques)

Les cas restants d'omissions (19,51 %) concernent des structures syntaxiques non décrites dans les grammaires de Colex. Les exemples relevés sont les phrases interrogatives, les structures avec verbes coordonnés (plus de deux verbes coordonnés) et les propositions infinitives ou gérondives dépendant d'un adjectif de contrôle (*The program is relatively easy to configure and use*). Ces constructions n'ont pas fait l'objet d'une description particulière à l'intérieur de Colex en raison de leur faible représentation dans le corpus d'étude.

L'analyse des erreurs de rappel présentée ci-dessus nous permet de souligner l'un des premiers défis rencontrés dans la construction du programme d'acquisition de collocations : concilier le mieux possible le double besoin de

1. représentativité des grammaires utilisées pour extraire les collocations dans les textes et

2. d'efficacité de l'algorithme de recherche.

Lorsque les besoins de rapidité d'exécution du programme (et d'autres impératifs de temps) nous ont amenée à limiter la couverture syntaxique de Colex aux constructions les plus courantes, nous avons cherché à compenser la perte de combinaisons lexicales qui pourrait en résulter en augmentant la taille du corpus. C'est ce qui a motivé l'inclusion de treize nouveaux documents (un total de 70 000 occurrences) au corpus d'origine.

Finalement, en mesurant le rappel (et la précision) de l'extraction des collocations, notre évaluation prend en charge les erreurs commises par

l'analyseur de Logos. Il s'agit là d'une conséquence de l'approche utilisée qui base l'extraction des collocations sur des connaissances linguistiques. L'analyse syntaxique des phrases par le logiciel de TA Logos est donc une composante essentielle de Colex. Les performances du programme d'acquisition de collocations sont ainsi directement liées aux performances de l'analyseur : une amélioration de la qualité de l'analyse syntaxique devrait entraîner une amélioration comparable de la qualité de l'extraction.

4.2.3.2 Mesure de la précision

Nous évaluons maintenant la précision du programme d'extraction de collocations, c'est-à-dire l'adéquation des résultats obtenus aux données recherchées. Pour évaluer la précision, nous avons examiné environ deux mille combinaisons (chaque occurrence des combinaisons extraites pour les termes COMPUTER, MEMORY et PROGRAM dans une des trois positions syntaxiques considérées). Rappelons que, pour l'instant, l'extraction des collocations se base sur des critères uniquement morphosyntaxiques. Les règles de Colex extraient des combinaisons verbe + nom décrivant une relation syntaxique (actancielle) particulière. Nous avons donc évalué la précision de l'extraction linguistique selon les mêmes critères et accepté, dans un premier temps, toute occurrence d'une combinaison qui remplissait les deux conditions suivantes :

1. la combinaison était formée d'un verbe et d'un nom et

2. le nom réalisait l'un des trois actants syntaxiques du verbe (sujet, premier complément, deuxième complément).

En appliquant ces deux critères à notre échantillon, nous obtenons un taux de précision de 85,15 % (281 occurrences incorrectes sur les 1 892 examinées). Le décompte des erreurs selon la combinaison verbe + nom extraite révèle des variations importantes entre les trois relations actancielles. Nous donnons ci-dessous le taux de précision de chaque relation :

Obj1 : 90,61 %

Subj : 94,16 %

Obj2 : 53,24 %

Les données ci-dessus font ressortir les faibles performances de Colex en ce qui concerne l'extraction de la troisième relation actancielle : un peu plus de la moitié seulement des combinaisons extraites pour cette relation sont correctes. L'analyse a permis de dégager huit types d'erreurs de précision décrits en détail dans les pages qui suivent. La distribution des erreurs de précision en fonction de chacune des trois relations actancielles est résumée dans le tableau ci-dessous.

	Subj	Obj1	Obj2	Total
ERR1	23	9	4	36
ERR2	4	13	2	19
ERR3	1	6	5	12
ERR4	0	3	69	72
ERR5	8	19	9	36
ERR6	0	3	61	64
ERR7	3	18	6	27
ERR8	0	12	3	15
Total	39	83	159	281

Tableau 4-II Répartition des différentes erreurs de précision selon le type de combinaison

La majorité des erreurs relevées dans notre échantillon (62,28 % des occurrences incorrectes) sont attribuables à des erreurs d'étiquetage et d'analyse des phrases d'où ont été extraites les combinaisons. Il s'agit en majeure partie d'occurrences uniques (une seule combinaison extraite du corpus d'étude). Nous examinons ci-dessous en détail les différentes erreurs relevées lors de l'évaluation de notre échantillon d'environ 2 000 occurrences.

Les deux premières catégories d'erreurs (ERR1-2) identifient des erreurs d'étiquetage de l'un ou l'autre membre d'une combinaison extraite. Toutes les erreurs relevées concernent l'élément verbal de la combinaison (19,57 % des erreurs de précision).

ERR1 – L'unité lexicale extraite de l'arbre comme gouverneur de la base nominale n'est pas un verbe. Il s'agit d'un homographe syntaxique qui n'a pas été correctement désambiguïsé, par exemple :

Major computer manufacturers even use computers to design and build better computers. ⇒ `*even(vt) + computer`

ERR2 – Erreur d'étiquetage sur le deuxième élément d'une locution verbale (verbe à particule). Le verbe extrait est incorrect :

Boot up the computer, and it launches Windows 95 and the TV software, and shifts into TV mode. ⇒ `*boot(vi)+up+computer`

Les trois catégories d'erreurs suivantes (ERR3-5) identifient des erreurs d'analyse par le logiciel de TA. Elles sont responsables du plus grand nombre de combinaisons fautives (42,70 % des combinaisons rejetées). Dans aucune de ces combinaisons, il n'existe de relation entre le verbe et le nom (ou groupe prépositionnel) extrait en tant que premier, deuxième ou troisième actant syntaxique. Nous avons dégagé les sous-catégories suivantes, en fonction du niveau d'analyse où l'erreur s'est produite.

ERR3 – Le nom extrait de l'arbre comme base nominale de la combinaison est un modificateur :

Some people hate computer cookies. → `*hate(vt) + computer`

ERR4 – Le nom (ou groupe prépositionnel) extrait de l'arbre comme base nominale de la combinaison est un complément de nom. Il s'agit de la plus importante catégorie d'erreurs (43,40 % des erreurs observées pour la relation verbe + deuxième complément) :

By consciously increasing your knowledge about your computer, you will learn about how you can take care of it. → `*increase(vt)+about+computer`

ERR5 – Le nom (ou le groupe prépositionnel) extrait de l'arbre comme base nominale de la combinaison est complément du verbe d'une autre proposition :

The first worldwide effort to open a global conversation using computers was over. → `*open(vt) + computer`

Les trois dernières catégories d'erreurs (ERR6-8) identifient des erreurs au niveau même de l'extraction effectuée par Colex : le programme

n'a pas analysé correctement la relation syntaxique entre le verbe et le nom (ou groupe prépositionnel) (37,72 % des erreurs).

ERR6 – Le groupe prépositionnel ne remplit pas le rôle de deuxième complément par rapport au verbe de la combinaison. Il s'agit d'un modificateur adverbial. Cette catégorie représente la deuxième plus importante catégorie d'erreurs (38,36 % des erreurs observées pour la relation verbe + deuxième complément) :

Graduate education programs need to provide young teachers with the information on teaching with computers, not just teaching about computers. ⇒ `*teach(vi)+with+computer`

ERR7 – Le nom (ou groupe prépositionnel) remplit un rôle différent de celui identifié par une des trois grammaires de Colex. Voir l'exemple ci-dessous d'une combinaison extraite pour la relation actancielle verbe + deuxième complément (il s'agit en fait d'une relation verbe + premier complément) :

We will start at the beginning with an extremely simple C program. ⇒ `*start(vi)+with+program`

ERR8 – Le nom ne remplit pas le rôle identifié par la grammaire d'acquisition par rapport au verbe de la combinaison. Il s'agit du sujet

d'un verbe d'état (ou de relation) au passif, extrait en position de premier complément. Rappelons que les règles d'acquisition de collocations à partir de structures passives rétablissent, pour l'extraction, l'ordre canonique des compléments — une méthode d'extraction qui inclurait des critères sémantiques à la découverte des collocations dans les textes pourrait facilement écarter ces combinaisons fautives de la liste des combinaisons retenues :

Semiconductor memory is composed of circuitry on silicon chips. ⟹ `*compose(vt) + memory`

4.2.3.3 Limites de l'approche morphosyntaxique au repérage des collocations

L'évaluation des combinaisons extraites pour les termes COMPUTER, MEMORY et PROGRAM à ce stade de l'extraction des collocations de la langue de l'informatique expose également la faiblesse principale d'une méthode d'extraction basée uniquement sur des critères morphosyntaxiques. En effet, si environ 85 % des combinaisons extraites pour les trois termes témoins illustrent bien les relations qui existent entre ces termes et les verbes de notre corpus, on ne saurait pour autant affirmer qu'il s'agit dans tous les cas de collocations. Ainsi, en ce qui concerne COMPUTER, nous avons relevé, à côté des verbes qui désignent les modalités d'utilisation spécifiques à ce terme *([to] boot, [to] connect, [to] disconnect, [to] interconnect, [to] network, [to] reboot, [to] restart, [to] shut off, [to] start, [to]*

turn off, *[to] turn on* et *[to] use*), un nombre tout aussi important de verbes avec lesquels l'unité lexicale est combinée librement (*[to] buy*, *[to] find*, *[to] get*, *[to] have*, *[to] need*, *[to] own*, *[to] ship back*, *[to] try out*, *[to] want*, ou même *[to] wear*). Nous retrouvons ces verbes parmi les cooccurrents extraits pour MEMORY (*[to] find*, *[to] get*, *[to] have*, *[to] need*, *[to] own*, *[to] ship*) et PROGRAM (*[to] acquire*, *[to] exchange*, *[to] get*, *[to] safeguard*, *[to] take*, *[to] try*).

Nous avons par ailleurs relevé un nombre important de combinaisons avec des verbes d'état et des verbes de relation qui définissent d'autres types de relations sémantiques entre les termes du domaine étudié (synonymie, hyperonymie, holonymie). Ainsi qu'il a été mentionné à propos des erreurs mettant en jeu des verbes d'état au passif (ERR8), l'accès à une caractérisation sémantique des verbes du corpus permettrait d'éliminer ces dernières combinaisons de la liste des combinaisons retenues. Nous donnons des exemples de ces combinaisons ci-dessous.

(33) *A computer is an electronic machine that enables a user to input, manipulate, store, and output information.*

(34) *A computer is made up of three basic parts: the system unit, the monitor, and the keyboard.*

(35) *Portable computers quite often have custom keyboards that have slightly different key arrangements than a standard keyboard.*

(36) *All computers have memory, also known as RAM.*

(37) *Laptop portable computers commonly include both a hard disk and a 3 ½-inch floppy diskette drive.*

Finalement, l'examen des combinaisons extraites par Colex reflète d'autres problèmes liés au manque d'interprétation sémantique des cooccurrents verbaux retenus pour une unité lexicale donnée. Dans un petit nombre de cas, le verbe extrait pour l'un des trois termes témoins est réellement contrôlé par l'unité lexicale qui remplit l'un des autres rôles prévus dans la structure actancielle de ce verbe. Voir par exemple les verbes *play* et *put* extraits des deux phrases suivantes comme collocatifs de COMPUTER — les collocations formées par ces deux verbes sont respectivement *[to] play a role* et *[to] put into use* :

(38) *Computers can also play a major role in improving the educational skills of our youth.*

(39) *When the computer is first put into use...*

Si nous incluons ces erreurs d'interprétation sémantique des combinaisons extraites aux erreurs syntaxiques relevées plus haut, nous constatons une importante diminution de la précision du programme d'acquisition. Elle passe ainsi de 85,15 % à 57,03 % : un peu plus de la moitié seulement des combinaisons extraites par Colex sont des collocations. Pour COMPUTER seulement, le terme le plus fréquent dans le corpus d'étude, la précision descend même à 47,62 % (407 occurrences libres ou incorrectes sur les 777 extraites pour ce terme). La précision varie également selon le type de combinaison extraite avec le taux de précision le plus faible pour la troisième relation actancielle :

Obj1 : 61,88 %;

Subj : 56,29 %;

Obj2 : 45,88 %.

Ces résultats démontrent clairement les limites d'une méthode d'extraction des collocations basée sur les seuls critères morphosyntaxiques. Ils ont motivé le développement et la mise en place du filtrage statistique décrit dans la troisième et dernière partie de ce chapitre.

4.3 Traitement statistique des sorties de Colex

4.3.1 Utilisation de la fréquence pour l'extraction des collocations

Nous abordons maintenant la dernière partie du chapitre consacré à la description du programme d'extraction de collocations développé dans le cadre de la présente recherche. Ainsi qu'il a été mentionné ci-dessus lors de l'évaluation de la précision du programme, un nombre très élevé de combinaisons incorrectes (83,73 %) sont des occurrences uniques. Le nombre de combinaisons libres ayant une seule occurrence est également élevé (62,50 % des combinaisons relevées). Parmi les combinaisons de deux occurrences, nous trouvons encore 10,05 % d'erreurs et 21 % de combinaisons libres.

Le traitement statistique présenté ci-dessous a pour objectif d'éliminer les erreurs et combinaisons libres contenues dans les fichiers résultats de Colex. Considérant la faible fréquence des combinaisons indésirables (une seule occurrence extraite dans la majorité des cas), nous avons d'abord envisagé de filtrer les combinaisons extraites par Colex

uniquement en fonction de leur fréquence (retenir les combinaisons qui apparaissaient le plus fréquemment dans le corpus d'étude). Une méthode de filtrage basée sur la fréquence absolue s'accordait avec les observations faites jusqu'à présent sur les phénomènes de cooccurrence lexicale, particulièrement le caractère répétitif de ces associations.

Nous avons également trouvé confirmation du bien-fondé d'une telle méthode dans les travaux d'autres chercheurs sur l'extraction automatique de combinaisons lexicales dans les textes. Dans la thèse qu'elle consacre à l'extraction automatique de terminologie, Daille (1994) constate ainsi que la fréquence permet le plus clairement d'isoler les phrasèmes terminologiques d'une liste de candidats termes après la mesure de liaison retenue en dernière analyse (le coefficient de vraisemblance). Dans une autre expérimentation (Krenn et Evert 2001), la courbe de la fréquence absolue des combinaisons verbe + groupe prépositionnel extraites de deux corpus de 8 et de 10 millions d'occurrences respectivement sert de référence pour les différentes mesures statistiques évaluées (information mutuelle, coefficient de Dice, test du chi-carré, test du rapport de vraisemblance, score T). Les auteurs concluent même qu'aucune de ces mesures classiques, à l'exception d'un test développé spécifiquement pour le type de combinaisons relevées, n'est mieux adaptée à l'identification des collocations dans la liste des combinaisons retenues que la seule fréquence de cooccurrence.

Nous sommes cependant obligée de rejeter l'utilisation de la seule fréquence pour filtrer les combinaisons extraites par Colex au vu d'un certain nombre d'observations que nous résumons maintenant. Nous constatons tout d'abord que, parmi les 20 premières combinaisons

extraites par Colex pour la relation verbe + premier complément, figure une majorité de combinaisons libres. Les combinaisons les plus fréquentes dans l'échantillon utilisé pour l'évaluation de Colex sont également des combinaisons libres : il s'agit pour les trois termes témoins des combinaisons formées avec les verbes *be*, *have* et *use*. Nous reproduisons ci-dessous la liste des vingt premières combinaisons (en termes de fréquence) pour la relation verbe + premier complément. Dans le reste de cette présentation, nous signalons les collocations véritables avec des caractères gras.

f:	101	be(vi)	one
f:	96	do(vt)	this
f:	93	see(vt)	figure
f:	70	use(vt)	it
f:	67	be(vi)	way
f:	**61**	**store(vt)**	**data**
f:	**59**	**use(vt)**	**command**
f:	58	be(vi)	program
f:	**49**	**send(vt)**	**data**
f:	46	be(vi)	file
f:	42	be(vi)	type
f:	40	be(vi)	system
f:	37	do(vt)	thing
f:	**37**	**be(vi)**	**part**
f:	36	do(vt)	it
f:	**35**	**take(vt)**	**look**
f:	35	be(vi)	version
f:	**34**	**use(vt)**	**computer**
f:	**34**	**run(vt)**	**program**
f:	**34**	**press(vt)**	**key**

Figure 4-6 Liste des vingt combinaisons les plus fréquentes pour la relation verbe + premier complément

Le deuxième problème que pose une méthode de filtrage basée sur la seule fréquence est que celle-ci ne nous permet pas de discriminer suffisamment les combinaisons retenues par Colex. En effet, un très grand

nombre de combinaisons ont la même fréquence. Cette deuxième observation concernant les combinaisons verbe + nom extraites par Colex est résumée dans le tableau ci-dessous.

Fréquence d'occurrence	*Nombre de combinaisons*	*Total des combinaisons*	*Nombre d'occurrences*	*Total des occurrences*
1	15 379	78,40%	15 379	48,62%
2	2 288	11,66%	4 576	14,47%
3	776	3,96%	2 328	7,36%
4	405	2,06%	1 620	5,12%
5	188	0,96%	940	2,97%
6	140	0,71%	840	2,66%
7	87	0,44%	609	1,93%
8	75	0,38%	600	1,90%
9	44	0,22%	396	1,25%
10	35	0,18%	350	1,11%
11	24	0,12%	264	0,83%
12	22	0,11%	264	0,83%
13	21	0,11%	273	0,86%
14	17	0,09%	238	0,75%
15	8	0,04%	120	0,38%
>15	≅3	0,55%	2 833	8,96%

Tableau 4-III Répartition des combinaisons verbe + premier complément en fonction de leur fréquence

Le tableau ci-dessus utilise deux valeurs pour résumer la distribution des combinaisons verbe + nom dans notre corpus d'étude. La première valeur (colonne 2) représente le nombre de combinaisons relevées avec une fréquence d'occurrence donnée ; pour les fréquences supérieures à 15, nous avons relevé en moyenne trois combinaisons. La deuxième valeur (colonne 4) représente le nombre total d'occurrences par catégorie de fréquence (calculé en multipliant le nombre de combinaisons de la colonne 2 par la fréquence).

Ainsi que l'illustre le Tableau 4-III, un très grand nombre de combinaisons (78,40 %) ont une fréquence 1 (apparaissent une seule fois dans le corpus). Ces hapax représentent toutefois moins de la moitié des occurrences effectivement extraites par Colex (48,62 % du total des occurrences). Le reste des occurrences extraites du corpus (plus de la moitié) correspond à des combinaisons reprises à un autre endroit du corpus (au moins deux occurrences relevées). Du point de vue du nombre d'occurrences extraites, la catégorie la plus importante après la catégorie de fréquence 2 est la catégorie des fréquences supérieures à 15 : un nombre important des occurrences extraites est associé à des combinaisons très fréquentes. Ces chiffres traduisent l'utilisation répétée d'un nombre important d'expressions linguistiques à travers le corpus d'étude, une des principales caractéristiques du discours spécialisé.

Le Tableau 4-III nous permet de faire une dernière constatation (qui invalide l'idée d'un filtrage des combinaisons basé sur la fréquence absolue) : parmi les combinaisons qui apparaissent plus d'une fois, 80 % environ apparaissent entre deux et quatre fois. Ces combinaisons ne seraient pas retenues par un filtre basé sur la fréquence absolue.

Nous avons donc décidé d'appliquer les méthodes de la statistique lexicale au problème du bruit dans les résultats de l'extraction linguistique des combinaisons verbe + nom de la langue de l'informatique. Comme il a été vu dans le deuxième chapitre, l'extraction des collocations dans les textes est un problème qui se prête particulièrement bien au traitement statistique. Les collocations décrivent les affinités lexicales spécifiques à une langue. Ces liens, généralement plus étroits que ceux des lexies

combinées librement, sont directement observables dans les textes et peuvent se mesurer.

Des chercheurs travaillant dans les domaines de la lexicographie et de la terminographie ont développé différentes mesures (ou adapté des tests statistiques utilisés dans d'autres domaines d'expertise) pour mesurer la force du lien entre les deux membres d'une combinaison lexicale donnée. Les mesures de liaison et tests statistiques développés (score T, test du chi-carré, test du rapport de vraisemblance, information mutuelle) comparent la probabilité d'observer les deux membres d'une combinaison ensemble (probabilité de la dépendance) avec celle de les observer séparément (probabilité de l'indépendance). Le rapport (la comparaison) de ces deux probabilités produit un score, appelé score d'association[6].

Dans la majorité des applications développées pour l'acquisition automatique de collocations, le score d'association n'est pas interprété, mais sert à ordonner les combinaisons évaluées, depuis celle présentant le meilleur score (ou l'association la plus forte) jusqu'à celle présentant le score le plus bas. Cette liste ordonnée de collocations candidates est ensuite soumise à un expert à des fins de validation.

Pour le filtrage des combinaisons extraites par Colex, nous avons choisi d'appliquer deux mesures d'association, l'information mutuelle et le

---

[6] Voir l'article de Kilgarriff (1996) pour un résumé des différentes mesures d'association utilisées dans l'analyse de corpus et le chapitre 5 de Manning et Schütze (1999) pour une description appliquée au repérage des collocations.

test du rapport de vraisemblance, et de comparer les résultats obtenus par les combinaisons acquises pour les dix termes les plus fréquents de notre corpus. Les deux mesures retenues pour l'évaluation figurent de façon marquée dans les travaux récents en matière d'acquisition automatique de collocations (Lin 1998; Kilgarriff et Tugwell 2001; Goldman *et al.* 2001). Dans tous ces travaux, elles sont intégrées à des applications qui basent l'extraction des collocations sur des critères morphosyntaxiques ; elles interviennent après l'extraction linguistique pour ordonner les combinaisons extraites en fonction du score obtenu par chacune (les combinaisons au plus fort potentiel collocationnel étant présentées en haut de la liste). Nous décrivons la mise en place du filtre statistique conçu dans le cadre de ce travail dans la section qui suit.

4.3.2 Description du programme ComputeScore

Le filtre statistique développé pour traiter les combinaisons verbe + nom extraites par Colex (ComputeScore) calcule le score d'association de chaque combinaison relevée selon l'une ou l'autre des deux mesures de liaison considérées, quelle que soit la fréquence de la combinaison. Dans l'examen des résultats obtenus avec les deux mesures, nous n'évaluerons cependant que les combinaisons ayant au moins deux occurrences. Deux facteurs ont influencé le choix de ce seuil de fréquence plutôt bas :

1. la taille du corpus (petite par rapport aux corpus utilisés dans des études similaires) et

2. le pré-traitement linguistique qui a permis d'éliminer une grosse partie du bruit dans les données brutes soumises habituellement au traitement statistique.

L'examen manuel des combinaisons extraites par Colex a par ailleurs révélé un nombre non négligeable de collocations parmi les combinaisons de fréquence 2 : *[to] carry information, [to] encrypt data, [to] recycle memory, [to] synchronize data, data flows, disk rotates, [to] host on server, [to] unload from memory*, etc. Nous avons donc jugé nécessaire de faire évaluer ces combinaisons par les filtres statistiques.

Les deux mesures de liaison utilisées, l'information mutuelle et le test du rapport de vraisemblance, calculent le score d'association d'une paire d'unités lexicales à partir de la fréquence de la paire (fréquence conjointe) et des fréquences des deux unités la composant (fréquences marginales). Ces différentes fréquences sont généralement résumées dans un tableau croisé, appelé également tableau de contingence[7] :

	N_j	$N_{j'}$ **avec** $j' \neq j$
V_i	a	b
$V_{i'}$ **avec** $i' \neq i$	c	d

Les variables a, b, c, et d représentent les fréquences suivantes :

a = fréquence d'une combinaison $V_i + N_j$

[7] Le tableau de contingence représenté ici et l'explication des variables qui suit sont empruntés à Daille (1994).

b = fréquences des combinaisons $V_i + N_{j'}$ ($N_{j'}$ représente un actant de même catégorie qui n'est pas N_j)

c = fréquences des combinaisons $V_{i'} + N_j$

d = fréquences des combinaisons $V_{i'} + N_{j'}$

Pour calculer le score d'association d'une combinaison $V_i + N_j$ donnée, les deux mesures utilisent également le total des fréquences des combinaisons extraites pour la relation actancielle représentée par $V_i + N_j$ (noté N). Nous donnons maintenant les formules des deux mesures de liaison, exprimées en fonction des valeurs du tableau de contingence.

Information mutuelle (*MI*) :

$$MI = \log_2 \frac{aN}{(a+b)(a+c)}$$

Test du rapport de vraisemblance (*Logl*) :

$$\begin{aligned} Logl = {} & a\log a + b\log b + c\log c + d\log d \\ & - (a+b)\log(a+b) - (a+c)\log(a+c) \\ & - (b+d)\log(b+d) - (c+d)\log(c+d) \\ & + N\log N \end{aligned}$$

Pour chaque combinaison $V_i + N_j$ extraite, Colex nous donne seulement la valeur de a (le nombre d'occurrences d'une combinaison $V_i + N_j$ donnée). Pour obtenir les valeurs correspondant à b, c et d dans le tableau de contingence ci-dessus (et calculer le score *MI* et *Logl* de $V_i + N_j$),

nous avons développé le programme indépendant ComputeScore[8]. ComputeScore prend le fichier résultat produit par Colex et calcule les valeurs des fréquences nécessaires au calcul du score d'association de chaque combinaison de ce fichier selon l'une ou l'autre des deux mesures de liaison considérées.

Le programme calcule en premier la valeur de *N*, le total des fréquences des combinaisons extraites pour une relation syntaxique donnée. Les acceptions intransitives et transitives des verbes extraits pour une relation actancielle donnée ont été rassemblés dans un seul fichier. Le programme calcule ensuite les valeurs de *b* et *c* (les fréquences marginales des unités lexicales composant chaque combinaison $V_i + N_j$ du fichier d'entrée). Il copie pour cela les combinaisons du fichier d'entrée dans une table de données. Chaque ligne de la table définit une paire $V_i + N_j$ particulière. Chacune des trois premières colonnes contient une des trois informations fournies par Colex pour chaque paire : fréquence, verbe (ou verbe+préposition), nom. La table de données générée à partir du fichier d'entrée de ComputeScore est représentée ci-dessous.

---

[8] ComputeScore a été programmé par mon amie Esmé Manandise à qui je dois aussi d'avoir enfin entrepris des études de linguistique. Pour cela et pour le reste, je lui redis ici toute ma reconnaissance.

f	*verbe*	*nom*
101	be(vi)	one
96	do(vt)	this
93	see(vt)	figure
70	use(vt)	it
67	be(vi)	way
61	store(vt)	data
59	use(vt)	command
58	be(vi)	program
49	send(vt)	data
46	be(vi)	file
42	be(vi)	type
40	be(vi)	system
37	do(vt)	thing
37	be(vi)	part
36	do(vt)	it
35	take(vt)	look
35	be(vi)	version
34	use(vt)	computer
34	run(vt)	program
34	press(vt)	key

Figure 4-7 Table de données générée à partir du fichier d'entrée de ComputeScore

Le programme parcourt ensuite la table et apparie les mots-formes à l'intérieur des deuxième et troisième colonnes dans la figure ci-dessus à chaque itération d'une boucle *while* pour fixer les valeurs de quatre autres variables pour chaque ligne $V_i + N_j$ de la table de données. Ces valeurs sont :

1. les fréquences des combinaisons où V_i apparaît comme premier élément — valeur inscrite dans la quatrième colonne de la table de données ;

2. le résultat de la soustraction de la fréquence de $V_i + N_j$ de la somme des fréquences de V_i — la différence correspond à la

valeur de b ; elle est inscrite dans la cinquième colonne de la table de données ;

3. les fréquences des combinaisons où N_j apparaît comme deuxième élément — valeur inscrite dans la sixième colonne de la table de données ;

4. le résultat de la soustraction de la fréquence de $V_i + N_j$ de la somme des fréquences de N_j — la différence correspond à la valeur de c ; elle est inscrite dans la septième colonne de la table de données.

Le programme calcule finalement le score *MI* de chaque paire $V_i + N_j$ de la table de données à partir des valeurs numériques des variables a, b et c de cette paire (colonnes 1, 5 et 7 de la table de données). Le score *MI* de la paire est inscrit dans la huitième (et dernière) colonne. Une autre sous-routine calcule le score *Logl*. Elle inclut le calcul de la valeur de d pour chaque paire de la table de données (les fréquences des combinaisons où ni V_i ni N_j n'apparaît).

Lorsque le score (*MI* ou *Logl*) de chaque paire de la table de données a été calculé, ComputeScore imprime les données de la table dans un fichier au format texte (MIScores.txt ou LoglScores.txt). Les combinaisons sont imprimées une par ligne et les données associées à chaque combinaison sont séparées par des points-virgules. Ces données texte sont importées dans un tableau Excel pour en faciliter le tri (ordonner les combinaisons selon le score obtenu, du score le plus élevé au score le plus bas). Nous présentons ci-dessous les vingt premières combinaisons verbe + premier complément selon la mesure d'information mutuelle.

a	verbe	nom	a + b	b	a + c	c	MI
2	spark(vt)	explosion	2	2	2	2	13,95
2	quicken(vi)	SE	2	2	2	2	13,95
2	button(vi) on	remote	2	2	2	2	13,95
2	finish(vi) from	last	2	2	3	1	13,36
2	recalculate(vt)	checksum	2	2	4	2	12,95
2	get(vi) from	jam	3	1	3	1	12,78
2	move(vi)	back and forth	3	1	3	1	12,78
2	balance(vt)	checkbook	5	3	2	2	12,63
2	weight(vt)	timing	5	3	2	2	12,63
2	hold(vt) from	kilobyte	2	2	5	3	12,63
2	welcome(vt)	contribution	4	2	3	1	12,36
2	heat(vt)	pass	3	1	4	2	12,36
2	absorb(vt)	red	4	2	3	1	12,36
2	deliver(vi) on	promise	2	2	6	4	12,36
2	gain(vi) in	popularity	2	2	6	4	12,36
2	arrive(vi) for	dick	2	2	6	4	12,36
2	state(vt)	intent	7	5	2	2	12,14
2	empty(vt)	recycle bin	7	5	2	2	12,14
2	direct(vi) by	telephone	2	2	7	5	12,14
2	subtract(vt)	tax	5	3	3	1	12,04

Figure 4-8 Vingt premières combinaisons verbe + premier complément selon l'information mutuelle

Ainsi que le montre la Figure 4-8, les paires qui obtiennent les scores les plus élevés selon l'information mutuelle ont la plus petite fréquence possible pour l'évaluation des filtres statistiques (2 occurrences seulement). La majorité d'entre elles sont formées d'unités lexicales qui apparaissent uniquement à l'intérieur de la paire : les fréquences de V_i et de N_j (colonnes 4 et 6 de la table ci-dessus) sont égales à la fréquence de $V_i + N_j$. Il s'agit de combinaisons incorrectes (*[to] button on remote, [to] finish from last*) ou non typiques pour le corpus considéré (*[to] spark explosion, [to] balance checkbook*). Ces résultats sont similaires à ceux rapportés par Daille (1994) lors de l'évaluation des performances de l'information mutuelle sur les combinaisons N_1 *de* (DET) N_2 extraites d'un corpus du

domaine des télécommunications. Ils soulignent le manque de fiabilité de l'information mutuelle lorsque le volume des données est faible — les fréquences recueillies pour l'une ou l'autre des deux unités lexicales sont trop petites. Ce phénomène est également rapporté dans les travaux récents sur l'extraction automatique de collocations qui, lorsqu'ils ne rejettent pas entièrement la mesure, ajustent celle-ci afin de compenser la surestimation des petites probabilités.

Nous donnons maintenant les vingt premières combinaisons retenues pour la relation verbe + premier complément selon le test du rapport de vraisemblance (les collocations sont indiquées en gras).

a	verbe	nom	a + b	b	a + c	c	Logl
93	see(vt)	figure	412	319	98	5	171,66
96	do(vt)	this	393	297	302	206	107,29
19	**keep(vt) in**	**mind**	19	19	28	9	86,12
16	**post(vi) to**	**newsgroup**	16	16	38	22	67,72
34	**press(vt)**	**key**	100	66	83	49	63,5
35	**take(vt)**	**look**	281	246	51	16	59,06
61	**store(vt)**	**data**	186	125	557	496	58,24
32	**press(vt)**	**button**	100	68	104	72	54,69
18	**pay(vt)**	**attention**	29	11	28	10	49,94
2	be(vi)	it	3754	3752	929	927	47,3
26	look(vi)	this	32	6	302	276	46,33
23	**play(vt)**	**game**	54	31	75	52	46,11
37	do(vt)	thing	393	356	128	91	38,3
27	**execute(vt)**	**instruction**	105	78	147	120	38,26
27	**take(vt)**	**advantage**	281	254	64	37	37,17
26	**do(vt)**	**job**	393	367	46	20	36,35
19	**solve(vt)**	**problem**	31	12	158	139	35,28
27	**send(vt)**	**email**	259	232	86	59	33,92
49	**send(vt)**	**data**	259	210	557	508	33,83
24	**perform(vt)**	**function**	124	100	126	102	32,33

Figure 4-9 Vingt premières combinaisons verbe + premier complément selon le test du rapport de vraisemblance

En contraste avec la liste des vingt premières paires selon l'information mutuelle, nous retrouvons ici quelques-unes des combinaisons qui figuraient parmi les vingt plus fréquentes dans le corpus (voir Figure 4-6). Ces données démontrent une plus grande corrélation entre le test du rapport de vraisemblance et la fréquence absolue. La liste ci-dessus présente toutefois des différences significatives avec celle des combinaisons les plus fréquentes. Elle intègre notamment beaucoup plus de collocations : parmi les vingt premières combinaisons d'après le test du rapport de vraisemblance, nous trouvons quinze collocations contre cinq seulement dans la liste basée sur la seule fréquence.

Nous allons maintenant examiner en détail les résultats obtenus pour les dix termes retenus pour l'évaluation des deux mesures d'association.

4.3.3 Évaluation des deux filtres statistiques

4.3.3.1 Élaboration d'une méthode d'évaluation

Pour évaluer les performances des deux filtres statistiques présentés ci-dessus (leur capacité à isoler les collocations dans la liste des combinaisons verbe + nom de la langue de l'informatique), nous avons cherché à utiliser une liste de référence (liste existante des collocations formées à partir des dix termes retenus pour l'évaluation de Colex). Cette liste devait nous permettre de contrôler l'adéquation des scores obtenus avec le caractère collocationnel d'une combinaison donnée.

Comme nous l'avons signalé dans le deuxième chapitre, les collocations verbe + nom (les collocatifs verbaux des termes) sont rarement répertoriées dans les terminologies. Il n'existe pas à notre connaissance d'ouvrage de référence, dans le domaine de l'informatique, qui pourrait servir à l'élaboration de la liste envisagée. L'un des premiers ouvrages consultés, Le *Dictionnaire d'Internet, de l'informatique et des télécommunications*, dictionnaire bilingue anglais-français publié par l'Office de la langue française, est essentiellement composé de noms ; cet ouvrage contient seulement quelques verbes, donnés avec leur traduction française uniquement.

Nous retrouvons cette approche à la description de la catégorie verbale dans le deuxième ouvrage de référence consulté, le *Dictionnaire d'informatique anglais-français* de Michel Ginguay — il s'agit également d'un dictionnaire bilingue anglais-français. De tous les ouvrages spécialisés consultés, ce dictionnaire est celui qui contient le plus grand nombre de verbes. Chaque verbe fait l'objet d'une entrée particulière qui donne sa traduction en français. Lorsque plusieurs traductions sont données — il s'agit généralement de traduire des sens collocationnels du verbe anglais — les différentes traductions sont différenciées à l'aide d'une indication contextuelle entre parenthèses (actant prototypique de l'acception désambiguïsée du verbe anglais). Nous donnons l'exemple de l'entrée de *[to] clear* :

> ***Clear (to)***, remettre à zéro (mémoire), effacer (écran), éteindre (voyant), vider (piste de cartes), supprimer (tabulations sur écran, bourrage), corriger (erreurs), habiliter (quelqu'un).

Le *Dictionnaire d'informatique* contient également un petit nombre de semi-phrasèmes verbaux (collocations verbe + nom). Suivant la pratique généralement adoptée dans les dictionnaires, ceux-ci sont donnés à l'entrée du verbe. Des exemples de collocations données par le *Dictionnaire d'informatique* incluent *[to] deny access*, *[to] draw a flowchart*, *[to] fail a test*, *[to] get on the computer* et *[to] place a call.*

Un seul dictionnaire se présente comme documentant les collocations de la langue de spécialité étudiée ici : il s'agit du *Répertoire*

bilingue de combinaisons lexicales spécialisées français-anglais d'Isabelle Meynard (2000). Il en a été question dans le deuxième chapitre. Ce dictionnaire énumère les collocations formées à partir de termes clés du domaine de l'Internet et contient nécessairement un nombre limité d'entrées (64 termes pour le français et l'anglais). Il s'agit d'une liste non exhaustive de combinaisons lexicales parmi les plus fréquentes dans un corpus de près de deux millions d'occurrences. Les combinaisons ont également été retenues sur la base d'une analyse contrastive entre les deux langues représentées : chaque combinaison retenue pour une langue avait un équivalent attesté dans l'autre langue. La vocation d'aide à la traduction de tous les ouvrages de référence trouvés pour le domaine de l'informatique explique en partie l'absence de description systématique des combinaisons verbe + nom, ces dernières n'étant données que pour expliciter le sens du verbe (anglais) spécialisé à l'intérieur de la combinaison et indiquer la traduction appropriée. La même motivation est à l'origine de TERMIUM, la banque de données terminologiques du Bureau de la traduction du gouvernement canadien. Elle justifie également que les collocations soient données essentiellement dans la partie française d'une fiche.

Face au manque de ressources dictionnairiques à l'aide desquelles valider les programmes d'acquisition automatique de collocations (particulièrement en ce qui concerne l'extraction de combinaisons verbe + nom), les chercheurs adoptent différentes solutions. Pour évaluer la précision avec laquelle les mesures de liaison identifient les collocations dans une liste d'environ 10 000 combinaisons verbales (verbe + groupe prépositionnel), Krenn et Evert (2001) utilisent une liste de référence

élaborée à partir des combinaisons ayant obtenu les scores les plus élevés selon l'une ou l'autre des mesures évaluées ; les combinaisons retenues ont ensuite été validées manuellement. Ces chercheurs sont les seuls, parmi ceux qui se sont intéressés à l'extraction de la cooccurrence verbale, à faire appel à une liste de référence. Dans les autres travaux consultés, les chercheurs évaluent la précision de l'extraction linguistique seulement (Lin 1998), font porter l'évaluation sur une autre caractéristique des combinaisons acquises (Goldman *et al.* 2001) ou laissent le lexicographe juger de la pertinence des combinaisons fournies par le programme d'extraction automatique (Kilgarriff et Tugwell 2001)[9].

Pour évaluer les performances des deux mesures de liaison considérées dans le cadre de notre recherche, nous avons élaboré un test similaire à celui décrit dans Krenn et Evert (2001). Dans un premier temps, nous avons déterminé manuellement les collocatifs verbaux de chaque terme témoin, dans chacune des trois positions syntaxiques considérées, à partir de la liste des combinaisons de fréquence supérieure ou égale à 2 extraites pour ce dernier.

Nous avons retenu les verbes qui illustraient les emplois prototypiques du terme considéré. Nous nous sommes aidée des exemples trouvés dans les ouvrages de référence mentionnés au début de cette section et avons ajouté aux listes de collocatifs ainsi constituées les verbes

---

[9] Dans les premières expériences d'acquisition automatique (Church et Hanks 1989; Smadja 1993), l'évaluation des listes était également faite par des lexicographes.

acquis automatiquement qui étaient sémantiquement reliés aux premiers : nous avons ainsi ajouté aux verbes *[to] load in* et *[to] load from* combinés avec MEMORY dans *Le grand dictionnaire terminologique* les deux verbes *[to] place in* et *[to] get from* extraits par Colex. (Cette méthode nous fait également retenir *[to] unload from.*)

Nous avons ensuite retenu les verbes qui exprimaient, pour les termes considérés, des valeurs de fonctions lexicales standard. Il s'agit le plus souvent de verbes supports ou de verbes de réalisation qui peuvent également exprimer le début, la fin ou la causation d'un événement. Nous donnons quelques exemples ci-dessous :

Pour PROGRAM, les verbes *[to] create, [to] write, [to] configure, [to] compile, [to] install, [to] open, [to] close*

Pour DISK, les verbes *[to] fill, [to] initialize, [to] read, [to] save on, [to] store on, [to] copy to, [to] write to, [to] load from, [to] read from, [to] transfer from*

Pour FILE, les verbes *[to] place in, [to] store in, [to] save to*

Finalement, nous avons retenu les verbes qui désignaient des actions très spécifiques, uniquement associées au terme donné, par exemple *[to] mirror a disk* ou *[to] encrypt data.* Il s'agit de sens spécialisés qui seront décrits à l'aide de FL non standard.

Avant de présenter les résultats de l'évaluation des deux mesures de liaison, nous aimerions signaler un autre aspect de cette évaluation : en évaluant les performances de chaque mesure à isoler les collocations dans

une liste de combinaisons candidates, nous n'avons pas cherché à établir de seuil en dessous duquel rejeter une collocation. Ainsi que le montreront les résultats obtenus, la détermination d'un tel seuil est difficile à faire, les données recueillies pour chacune des trois relations actancielles étant loin d'être homogènes. Ainsi, pour l'unité lexicale FILE, les combinaisons obtenant les plus hauts scores pour les trois types de combinaisons extraites sont respectivement *file contains* (16,34), *[to] copy file* (20,29) et *[to] save to file* (5,74). Des résultats similaires sont observables à l'intérieur d'une même relation actancielle : la combinaison la plus forte pour DATA en position d'objet direct (*[to] store data*) reçoit un score de 58,24 ; la plus forte pour SOFTWARE (*[to] install software*) a un score de 6,86.

Nous avons donc suivi la pratique adoptée dans tous les outils d'extraction de collocations considérés (qui utilisent seulement le score pour ordonner les combinaisons) et substitué à la notion de seuil de signification celle de liste de signification.

4.3.3.2 Résultats de l'évaluation des deux mesures de liaison

Une fois dressée la liste des collocatifs des dix termes témoins pour les trois relations actancielles, nous faisons porter l'évaluation des deux mesures de liaison sur les listes des combinaisons ordonnées selon l'une ou l'autre mesure (un total de 715 combinaisons). En partant du haut de chacune des deux listes évaluées (de la combinaison ayant obtenu le meilleur score), nous mesurons la proportion de collocations (la précision) dans des sous-ensembles de plus en plus larges de cette liste et traçons la courbe de la précision de la mesure correspondante avec les pourcentages obtenus pour chaque sous-ensemble. Les courbes obtenues pour les deux

mesures de liaison évaluées sont présentées dans le graphique ci-dessous. L'axe des ordonnées donne le pourcentage de collocations et l'axe des abscisses le sous-ensemble correspondant de la liste de combinaisons évaluées.

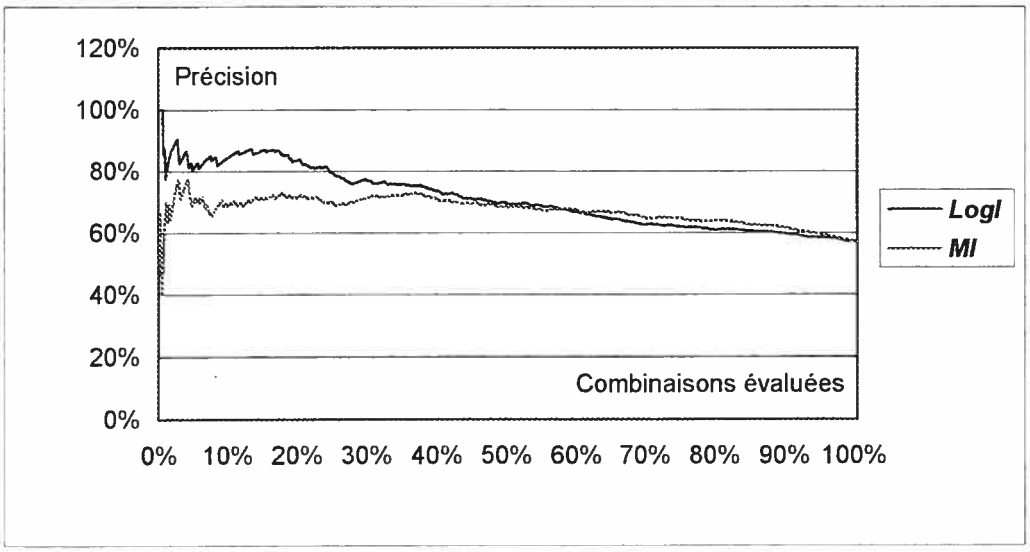


Figure 4-10 Courbes de précision pour les dix termes témoins

Ainsi que le montre le graphique ci-dessus, la courbe de la précision du test du rapport de vraisemblance (*Logl*) est clairement meilleure que celle de l'information mutuelle (*MI*) : les deux premiers tiers de la liste des combinaisons ordonnées selon *Logl* (60 % environ) intègrent beaucoup plus de collocations (77 % en moyenne) que les deux tiers correspondants de la liste *MI*.

Si nous comparons maintenant les résultats obtenus par l'une ou l'autre des deux mesures de liaison avec les listes des combinaisons extraites pour chaque relation actancielle (listes **Subj**, **Obj1** ou **Obj2**), nous

constatons des différences moins marquées entre les deux mesures. Pour les listes ordonnées des combinaisons verbe + premier complément (411 combinaisons), la courbe *Logl* est nettement inférieure à la courbe *MI* en ce qui concerne les premières combinaisons des listes : deux combinaisons libres, *[to] tell server* et *[to] be data*, reçoivent un score très élevé selon le test du rapport de vraisemblance et font ainsi chuter la courbe de précision de la mesure. (Celle-ci reste cependant à près de 80 %.) Par la suite, la courbe *Logl* est de nouveau supérieure à la courbe *MI* jusqu'au début du dernier tiers des listes de combinaisons ordonnées où les deux courbes se rejoignent.

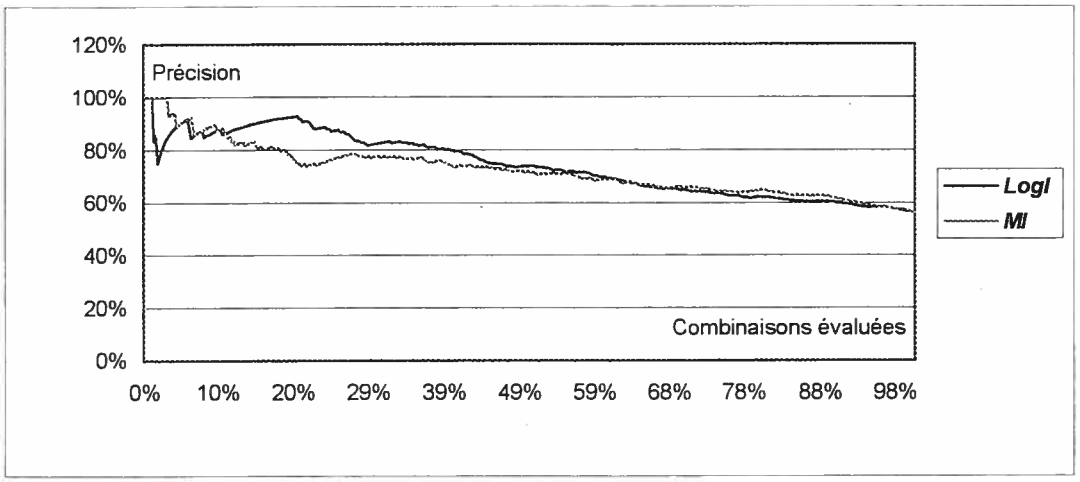


Figure 4-11 Courbes de précision pour les combinaisons verbe + premier complément

Nous examinons maintenant en détail les résultats obtenus pour la relation verbe + premier complément avec les combinaisons extraites pour FILE, l'unité lexicale pour laquelle Colex a extrait le plus grand nombre de combinaisons. Le tableau qui suit énumère les vingt-cinq premiers

cooccurrents de FILE pour la deuxième relation actancielle selon *Logl* (à gauche) et *MI* (à droite). Les collocatifs qui représentent des valeurs de fonctions lexicales standard sont en caractères gras. Les collocatifs non standard sont en caractères maigres sans italiques. Les combinaisons libres sont en italiques maigres.

Logl	*f*	*verb*	*MI*	*f*	*verb*
20,29	21	**copy(vt)**	5,75	2	tune(vt)
13,23	18	**download(vt)**	5,09	7	**upload(vt)**
12,89	20	**open(vt)**	4,92	9	parse(vt)
12,32	21	**save(vt)**	4,33	12	**back up(vt)**
11,78	12	**back up(vt)**	4,28	21	**copy(vt)**
10,89	9	parse(vt)	4,16	2	*designate(vt)*
9,03	7	**upload(vt)**	4,16	4	*double-click(vi) on*
5,39	9	**delete(vt)**	4,16	3	rename(vt)
4,84	7	**name(vt)**	4,01	3	exchange(vt)
4,81	9	locate(vt)	3,94	2	overwrite(vt)
4,1	10	move(vt)	3,75	2	write(vi) to
4,08	2	tune(vt)	3,63	3	**erase(vt)**
3,8	7	attach(vt)	3,59	18	**download(vt)**
3,8	8	*have(vt)*	3,58	2	*swap(vt)*
3,67	4	*double-click(vi) on*	3,47	7	**name(vt)**
3,39	6	**update(vt)**	3,30	20	**open(vt)**
2,99	5	**modify(vt)**	3,16	9	**delete(vt)**
2,79	5	*highlight(vt)*	3,16	5	**modify(vt)**
2,75	3	rename(vt)	3,11	21	**save(vt)**
2,65	17	**create(vt)**	3,05	6	**update(vt)**
2,6	3	exchange(vt)	3,02	5	*highlight(vt)*
2,29	46	*be(vi)*	3,01	3	*search(vi) for*
2,22	3	**erase(vt)**	2,97	7	attach(vt)
2,13	13	send(vt)	2,94	3	*drag(vt)*
2,09	8	access(vt)	2,94	2	*help(vt)*

Tableau 4-IV Comparaison des vingt-cinq premiers verbes retenus pour FILE

L'examen des vingt-cinq premiers verbes retenus pour FILE par l'une et l'autre des mesures de liaison considérées ici révèle des différences importantes entre les deux mesures. Parmi les vingt-cinq premières combinaisons de FILE, *Logl* inclut quatre combinaisons libres alors que *MI*

en inclut sept. Douze collocatifs sur les vingt et un retenus par *Logl* représentent des valeurs de fonctions lexicales standard : ils expriment les actions prototypiques associées à FILE. Ces verbes reçoivent également les scores les plus élevés selon cette mesure. *MI* retient onze de ces verbes et répartit la majorité d'entre eux dans la deuxième moitié de la liste ; la mesure leur accorde moins d'importance.

Le classement des collocatifs verbaux de FILE offert par *Logl* est donc meilleur que celui offert par *MI* : il permet une meilleure appréhension de la combinatoire verbale de cette unité lexicale. La tendance déjà relevée de l'information mutuelle à privilégier les occurrences rares explique ici encore les moins bonnes performances de la mesure : elle privilégie un plus grand nombre de verbes non caractéristiques tels que *[to] tune* (donné comme le plus significatif pour cette unité lexicale), *[to] designate*, *[to] swap*, *[to] search for* ou *[to] drag*.

Nous présentons maintenant les résultats obtenus par l'information mutuelle et le test du rapport de vraisemblance avec les listes des combinaisons extraites pour la première et la troisième relation actancielle.

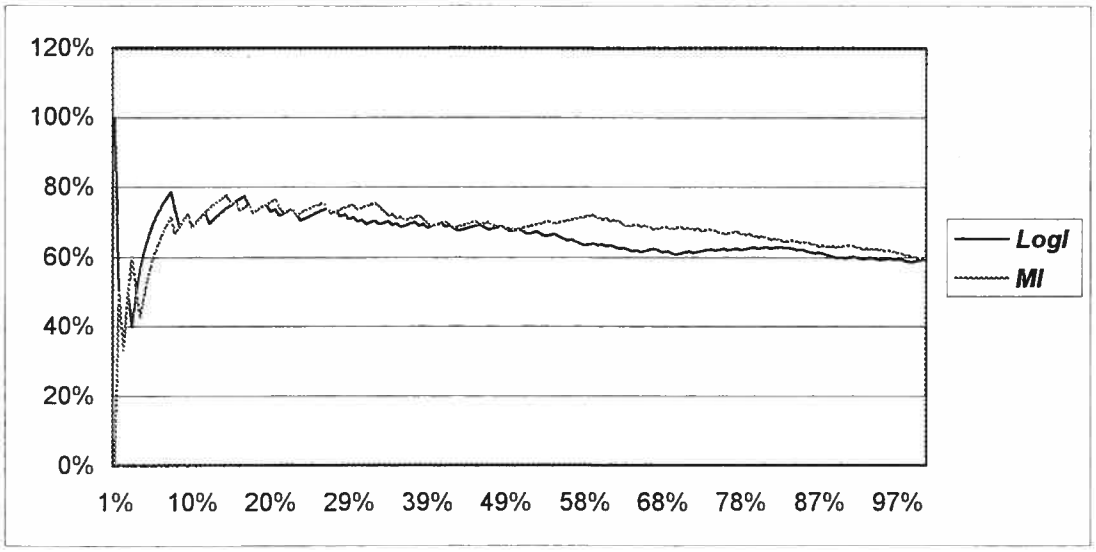


Figure 4-12 Courbes de précision pour les combinaisons sujet + verbe

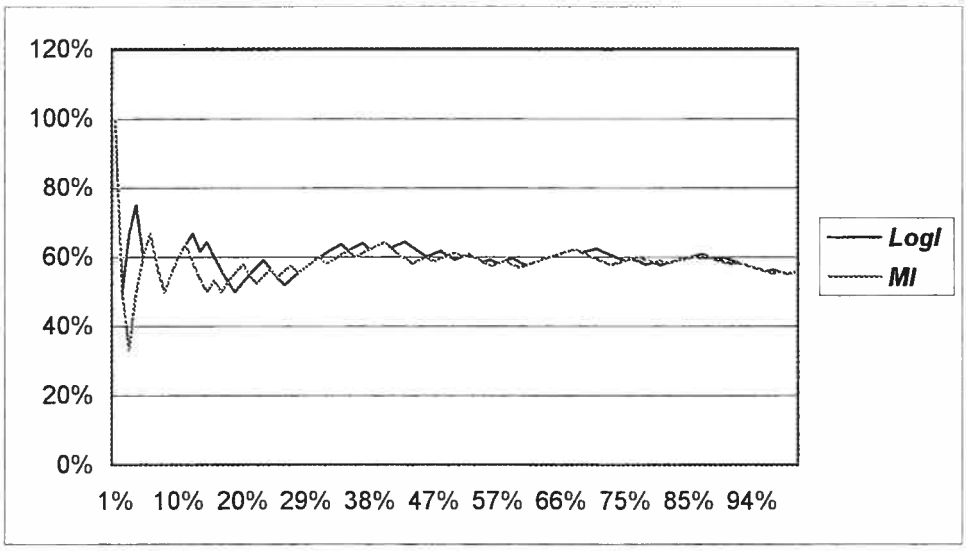


Figure 4-13 Courbes de précision pour les combinaisons verbe + deuxième complément

Ainsi que l'illustrent les deux graphiques ci-dessus, *Logl* et *MI* obtiennent pratiquement les mêmes résultats avec les combinaisons extraites pour la première et la troisième relation actancielle, *Logl* affichant de meilleurs résultats seulement pour les vingt premières combinaisons des listes **Subj** (**Obj2**) évaluées. Pour le reste des combinaisons, la précision des deux mesures est plus ou moins constante : elle se maintient autour de 60 % pour les combinaisons verbe + deuxième complément, traduisant une distribution plus uniforme des collocations à l'intérieur de chaque liste ordonnée.

On peut trouver une explication possible à la faible précision de *Logl* et de *MI* en ce qui concerne les combinaisons acquises pour la première et la troisième relation actancielle dans le manque de données disponibles pour l'évaluation. Pour la relation sujet + verbe, 194 combinaisons seulement ont été évaluées. Elles n'étaient que 110 pour la relation verbe + deuxième complément. Les combinaisons acquises pour la première et la troisième relation actancielle étaient également moins fréquentes : seules 15 combinaisons parmi les 194 acquises pour la relation sujet + verbe avaient une fréquence d'occurrence supérieure à 10. Cette proportion était encore plus réduite pour les combinaisons verbe + deuxième complément puisque deux combinaisons seulement sur 110 avaient une fréquence d'occurrence supérieure à 10.

Nous concluons donc cette évaluation des deux filtres statistiques en retenant le test du rapport de vraisemblance comme la mesure d'association la plus apte à isoler les collocations parmi les combinaisons extraites par Colex : la précision de *Logl* pour l'ensemble des combinaisons évaluées ci-dessus est de 71 %, elle est de 67 % pour *MI*. Comme l'a

montré l'évaluation des combinaisons extraites pour la première et la troisième relation actancielle, le filtrage statistique pose cependant certains problèmes. Les meilleurs résultats obtenus lors de l'évaluation des combinaisons verbe + premier complément nous permet toutefois d'espérer une amélioration de la précision du filtre statistique proportionnelle au volume des données traitées (particulièrement en ce qui concerne les combinaisons sujet + verbe et verbe + deuxième complément).

Nous avons terminé la présentation du programme d'acquisition de collocations développé dans la présente thèse. Dans le chapitre qui suit, nous résumons les principaux aspects de ce programme. Nous y présentons également quelques solutions possibles aux problèmes soulevés par l'extraction automatique des collocations de la langue de spécialité.

5 Conclusion

Le but de la présente recherche était de construire un modèle d'extraction des collocations de la langue de spécialité. En développant ce modèle, nous voulions également formaliser la notion de collocation, et particulièrement de collocation verbale, en terminologie. Afin d'extraire automatiquement les collocatifs verbaux des termes, nous avons :

1. analysé un corpus de textes de l'informatique de 600 000 mots-occurrences ;

2. développé des règles qui extraient les combinaisons formées du verbe et de l'un de ses actants syntaxiques des analyses des phrases du corpus ;

3. mis en place un filtre statistique sous la forme d'un score d'association pour isoler les collocations des combinaisons extraites.

Nous reprenons les trois aspects de la méthode développée ci-dessous. Pour chacun, nous soulignons notre contribution à l'étude des collocations en langue de spécialité. Les limites et les perspectives de notre travail sont également présentées.

5.1 Analyse du corpus

La méthode d'extraction développée dans le cadre de la présente recherche est basée sur un corpus analysé : les collocations sont extraites d'arbres qui représentent la structure syntaxique des phrases du corpus.

Notre méthode s'inscrit ainsi dans la continuation des travaux les plus récents en matière d'extraction automatique de collocations (Lin 1998; Kilgarriff et Tugwell 2001; Goldman *et al.* 2001). Elle se distingue d'autres méthodes qui basent l'extraction des collocations sur des corpus analysés par le modèle de description des collocations qu'elle implémente — celui des fonctions lexicales de la théorie Sens-Texte. Nous avons montré à la section 2.1.4 l'intérêt de la modélisation des collocations au moyen des fonctions lexicales par rapport à d'autres modélisations proposées.

Les fonctions lexicales sont un outil de description formel qui permet une caractérisation syntaxique et sémantique des phénomènes collocationnels d'une langue. Les collocations verbales qui nous ont particulièrement intéressée ici représentent ainsi les trois relations syntaxiques majeures, entre le verbe et le sujet grammatical, le complément d'objet direct et le complément d'objet indirect.

Nous avons donc basé l'extraction des collocations sur les analyses, par le système de TA Logos, des phrases d'un corpus informatique anglais. Il s'agit d'analyses complètes des phrases qui identifient également les compléments extraposés des verbes, nous permettant d'obtenir un échantillon plus large de combinaisons verbales.

L'analyse syntaxique effectuée par le système Logos repose sur SAL, l'ensemble d'étiquettes sémantico-syntaxiques qui permet de représenter les unités lexicales de l'anglais dans le dictionnaire du système. Au début du traitement automatique, chaque unité lexicale de la phrase est remplacée par un ensemble de valeurs SAL qui spécifient la partie du discours et les caractéristiques sémantico-syntaxiques et

morphosyntaxiques de l'unité. L'étiquetage sémantico-syntaxique de Logos a servi de base au formalisme que nous avons développé pour représenter, dans des règles d'extraction, les structures morphosyntaxiques caractéristiques des collocations verbales.

5.2 Développement des règles d'extraction

Les règles développées pour extraire les collocations des arbres produits par l'analyseur de Logos modélisent les trois relations actancielles représentées par les fonctions lexicales verbales : le terme remplit le rôle de sujet, premier complément ou deuxième complément du verbe. Les règles d'extraction sont regroupées dans trois grammaires spécialisées dans l'extraction d'une relation actancielle particulière. Notre programme se distingue ainsi de la plupart des programmes utilisés pour l'extraction automatique des collocations qui extraient les seules relations sujet + verbe et verbe + objet.

Les règles d'extraction identifient les structures syntaxiques de surface caractéristiques des collocations verbales et extraient les combinaisons formées du verbe et de son sujet, premier ou deuxième complément de l'arbre correspondant. Elles extraient donc des paires d'unités lexicales. En extrayant les combinaisons formées du verbe et de l'un de ses actants, les règles rétablissent l'ordre canonique des compléments et calculent également la fréquence (le nombre d'occurrences) de chaque combinaison extraite.

Les grammaires d'extraction sont intégrées dans un programme (Colex) qui contrôle l'appariement de chaque règle avec les arbres du

corpus et exécute les instructions qu'elle contient : extraction des unités lexicales associées aux nœuds représentant le verbe et son premier, deuxième ou troisième actant syntaxique et calcul de la fréquence de la combinaison verbe + terme ainsi formée. Le programme extrait une relation actancielle à la fois. À chaque exécution du programme, les combinaisons illustrant une relation syntaxique donnée sont réparties dans deux fichiers différents, en fonction de l'acception, transitive ou intransitive, du verbe. Les combinaisons sont données dans chaque fichier avec leur fréquence d'occurrence.

Pour évaluer la méthode d'extraction développée dans le cadre de la présente recherche, nous avons tout d'abord mesuré le rappel de l'extraction syntaxique des combinaisons verbe + terme. Sur les 213 combinaisons sélectionnées au hasard pour l'évaluation du rappel, 131 seulement ont été extraites automatiquement, ce qui représente un rappel de 61,50 %. Ce faible taux reflète en grande partie les limites de l'analyseur utilisé. Un pourcentage moins élevé d'omissions (20 % environ) correspondent à des structures syntaxiques qui ne sont pas décrites dans les grammaires de Colex.

Nous avons en effet cherché à limiter la taille des grammaires pour des raisons d'efficacité : dans l'appariement des structures syntaxiques décrites dans les règles d'une grammaire, l'ensemble des règles est appliqué à chaque arbre du corpus. Pour améliorer le rappel de Colex, nous pourrions optimaliser l'algorithme d'appariement du programme et développer la couverture des grammaires d'extraction.

Nous avons également mesuré la précision de l'extraction syntaxique des combinaisons verbe + terme. Elle est beaucoup plus élevée que le rappel. Environ 85 % des combinaisons extraites par Colex représentent une des trois relations actancielles verbe + sujet, verbe + premier complément ou verbe + deuxième complément. (L'évaluation a porté sur un échantillon de 2 000 combinaisons.) Nous avons cependant constaté des variations importantes entre les trois relations actancielles en ce qui concerne la précision. Ainsi, la précision de la troisième relation actancielle (verbe + deuxième complément) est de 53 % environ seulement. La faible précision de l'extraction syntaxique de ces combinaisons est attribuable à une mauvaise analyse du groupe prépositionnel extrait en tant que deuxième complément : il représente, dans un peu moins de la moitié des cas, un modificateur adverbial du verbe.

Une solution pour améliorer la précision de l'extraction des deuxièmes compléments consisterait à extraire dans cette position les seuls groupes prépositionnels analysés par Logos comme des actants syntaxiques du verbe. Nous avons cependant écarté cette solution au problème du rôle syntaxique des compléments prépositionnels, le régime du verbe anglais n'étant pas décrit systématiquement dans le dictionnaire du système. Nous pourrions également implémenter dans l'outil une stratégie pour distinguer automatiquement les deux types de complément du verbe[1].

---

[1] Une telle stratégie est développée dans Fabre et Frérot (2002). Elle serait facilement implémentable dans Colex.

Le taux de précision présenté ci-dessus mesure seulement l'extraction syntaxique des combinaisons verbales. Lorsqu'on ajoute des critères sémantiques à l'évaluation, la précision du programme tombe à 57,03 %. Cette faible précision reflète les limites d'une méthode d'extraction qui ne considère que les propriétés syntaxiques des combinaisons extraites. Or les collocations ne décrivent pas seulement les relations syntaxiques qui existent entre le verbe et ses actants. Elles expriment également des relations sémantiques particulières qui correspondent à un petit ensemble de significations prototypiques et manifestent des automatismes dans la construction du discours.

Les moins bonnes performances de Colex lorsque l'on considère également le sens des combinaisons extraites reflète le fait qu'un grand nombre de ces combinaisons représentent des combinaisons libres (combinaisons formées uniquement selon leur sens individuel). Ces combinaisons font chuter la précision de l'extraction, la plus forte baisse étant enregistrée pour la troisième relation actancielle : seulement 46 % environ des combinaisons extraites pour cette relation représentent des collocations véritables.

Afin d'améliorer la précision de notre méthode d'extraction, nous avons développé un filtrage statistique des combinaisons extraites. Une autre solution est également envisageable. Elle consisterait à encoder les combinaisons retenues sous la forme d'une fonction lexicale. Seules les combinaisons pour lesquelles un tel encodage est possible seraient alors retenues comme des collocations véritables.

5.3 Mise en place du filtre statistique

Le filtre statistique a pour but d'éliminer les combinaisons libres des fichiers résultats de Colex. Le filtrage statistique s'appuie sur les propriétés sémantiques des collocations et sur le caractère essentiellement contraint de leur mode de composition. Ce caractère se manifeste par une distribution particulière des unités lexicales ainsi associées dans les textes et peut donc se mesurer.

Différentes mesures statistiques ont été développées en linguistique de corpus, pour le travail lexicographique et terminographique, pour évaluer la force des associations lexicales d'un texte (leur caractère contraint). Dans la majorité des cas, les mesures produisent un score qui est ensuite utilisé pour ordonner les combinaisons.

Parmi les mesures développées, deux en particulier figurent de façon marquée dans les travaux récents en matière d'acquisition automatique de collocations. Il s'agit de l'information mutuelle (*MI*) et du coefficient de vraisemblance (*Logl*). Nous avons donc choisi d'évaluer l'efficacité de ces deux mesures à isoler les collocations d'un texte et développé un programme (ComputeScore) qui calcule le score d'association des combinaisons extraites par Colex selon les deux mesures évaluées. ComputeScore produit deux listes pour chacune des deux mesures évaluées. Les combinaisons sont ordonnées à l'intérieur de chaque liste selon leur score (*MI* ou *Logl*).

L'évaluation de la précision des deux mesures d'évaluation les plus communément utilisées pour isoler les combinaisons significatives d'un

texte représente un aspect original de notre recherche. En effet, bien qu'elles soient implémentées dans tous les programmes d'extraction automatique de collocations, les mesures d'association ne font généralement l'objet d'aucune évaluation rigoureuse. La seule évaluation du filtrage statistique de combinaisons extraites sur des bases morphosyntaxiques se trouve dans Krenn et Evert (2001). Le manque d'ouvrages de référence à l'aide desquels valider les résultats du filtrage statistique explique cette presque totale absence d'évaluations portant spécifiquement sur celui-ci. Le problème est particulièrement important en langue de spécialité. Afin d'évaluer les performances de différents filtres statistiques à retenir les collocations dans une liste de combinaisons verbales, Krenn et Evert utilisent une liste de contrôle élaborée à partir des combinaisons ayant obtenu les meilleurs scores selon chacune des mesures évaluées.

Pour évaluer les deux filtres statistiques, nous avons nous aussi élaboré une liste de contrôle à partir des verbes retenus pour les dix termes les plus fréquents de notre corpus. Afin d'arrêter la liste des collocatifs verbaux de nos termes, nous avons appliqué des critères sémantiques basés sur les fonctions lexicales : le verbe illustrait un emploi prototypique du terme — il s'agissait par exemple d'un verbe de réalisation, typique des artefacts.

Notre évaluation des performances du filtrage statistique s'est donc faite en fonction de critères plus contraignants que ceux utilisés par Krenn et Evert. Le filtre sélectionné devait retenir en priorité les verbes qui exprimaient des valeurs de fonctions lexicales standard. Nous retenons en priorité ces cooccurrents verbaux des termes parce qu'ils correspondent à

des connaissances lexicales véritables (devant être répertoriées). L'étude d'un nombre suffisamment représentatif de collocations potentielles devrait également nous permettre de développer des heuristiques d'encodage de ces liens lexicaux.

Nous avons comparé les deux listes ordonnées des combinaisons extraites pour les trois relations actancielles à notre liste de contrôle. Nous voulions dans un premier temps déterminer quelle liste ordonnée retenait le plus de collocations dans la moitié supérieure de la liste. La liste ordonnée selon le coefficient de vraisemblance présente clairement les meilleurs résultats : 79 % en moyenne des combinaisons classées dans la moitié supérieure de la liste ordonnée selon cette mesure sont des collocations (*MI* retient 70 % seulement des collocations). La précision de *Logl* pour l'ensemble de la liste est de 71 %. Nous constatons de nouveau des variations entre les trois relations actancielles en ce qui concerne la précision : alors qu'elle atteint 74 % pour les combinaisons verbe + premier complément, elle n'est que de 65 % pour les combinaisons sujet + verbe et descend à 59 % pour les combinaisons verbe + deuxième complément.

Cette moins bonne précision du filtre statistique pour ces deux types de combinaisons, qui traduit une répartition plus uniforme des collocations à l'intérieur des deux listes de combinaisons ordonnées, peut s'expliquer de la façon suivante : nous avions moins de données (combinaisons) pour ces deux relations et les combinaisons extraites étaient également moins fréquentes. Le coefficient de vraisemblance est corrélé avec la fréquence absolue d'une combinaison. Une première solution à ce problème consisterait à augmenter la taille du corpus.

Une deuxième solution au problème du manque de précision de la mesure retenue pour isoler les collocations d'une liste de combinaisons verbales consisterait à essayer d'encoder chaque combinaison sous la forme d'une fonction lexicale standard. Les fonctions lexicales modélisent les collocations. L'aptitude d'une combinaison donnée à s'encoder comme une fonction lexicale standard pourrait donc servir de base au développement d'un filtre sémantique des combinaisons extraites par Colex.

Ainsi qu'il a été présenté dans l'état de l'art, les fonctions lexicales syntagmatiques encodent deux types d'information :

1. le contenu sémantique et

2. la structure syntaxique des combinaisons lexicales décrites.

Pour encoder automatiquement les combinaisons extraites par Colex, il importe de dissocier formellement les deux types d'information et de proposer deux formules distinctes pour représenter l'information symbolisée par chaque fonction lexicale[2].

Il s'agirait dans un premier temps de faire apparaître explicitement les structures syntaxiques des fonctions lexicales verbales afin de déterminer les structures pouvant décrire une collocation candidate, par

---

[2] Une explicitation de la structure syntaxique des fonctions lexicales supports est donnée dans Heid (1996). Plus récemment, Kahane et Polguère (2001) décrivent deux formalismes d'encodage qui explicitent la structure et le sens des fonctions lexicales.

exemple une combinaison sujet + verbe pourra s'encoder comme **Func$_0$** ou **Func$_i$**, une combinaison verbe + premier complément comme **Oper$_i$** mais aussi **Caus** (cette fonction représente un verbe causatif). Une combinaison sujet + verbe ne peut jamais s'encoder comme **Caus**, la position du premier actant syntaxique étant réservée pour le causateur.

Certaines fonctions lexicales verbales, notamment les fonctions exprimant les trois phases d'un événement (**Incep**, **Cont**, **Fin**), n'ont pas de structure syntaxique associée. Dans ce cas, l'encodage devra s'appuyer sur la seule description sémantique de la fonction lexicale.

Il est possible d'établir un premier niveau d'encodage des combinaisons extraites en fonction de la structure syntaxique qu'elles décrivent, par exemple

Pour la combinaison *program runs* : **Func$_0$**, **Func$_i$**, **Fact$_0$**, **Fact$_i$**, **Incep**, etc.

Pour la combinaison *[to] launch a program* : **Oper$_i$**, **Real$_i$**, **Caus**, **Incep**, etc.

L'encodage du sens d'une combinaison particulière (la sélection d'une fonction lexicale parmi celles qui décrivent la même relation syntaxique) représente la partie la plus dure de l'encodage automatique.

Trente trois fonctions lexicales existent pour représenter les sens des collocatifs verbaux des termes, permettant une description sémantique très fine. Elles peuvent être combinées et accroissent ainsi la portée descriptive du modèle. La richesse de ces moyens descriptifs est justifiée

par le nombre et la variété des collocations. Nous redonnons ci-dessous des exemples de combinaisons verbe + premier complément pour PROGRAM (selon leur ordre d'importance d'après le coefficient de vraisemblance) : *[to] run, [to] execute, [to] invoke, [to] write, [to] compile, [to] exit, [to] start, [to] download, [to] call, [to] install, [to] close, [to] activate, [to] load, [to] create, [to] open, [to] configure, [to] launch, [to] remove a program.*

Les fonctions lexicales encodent des sens aussi variés que ceux donnés en exemple ci-dessus. Chaque fonction définit un ensemble homogène de combinaisons d'unités lexicales, reliées par la même relation lexicale. On peut les décrire comme des métalexies (cf. Polguère 2003b) dont le sens ne peut être appréhendé qu'en pratique, dans l'application de la fonction à une unité lexicale donnée. Parce qu'il s'agit d'une unité lexicale « générale », le sens d'une fonction lexicale est nécessairement vague et ne peut être véritablement décrit (calculé) que sur la base d'un échantillon représentatif de collocations types.

C'est l'approche adoptée par Wanner et Alonso Ramos (2001) à la description du contenu sémantique d'une fonction lexicale. La méthode développée dérive le sens d'une fonction lexicale (**Oper$_1$**) d'une collection de collocations référentielle. Le sens de la fonction lexicale est établi à partir de combinaisons possibles d'hyperonymes (ceux donnés dans WordNet pour chaque base et chaque collocatif de la collection référentielle) et représenté par un indice : la pertinence moyenne des collocations de la collection référentielle.

Cet indice sert ensuite à évaluer les nouvelles collocations (en comparant les résultats obtenus par la collocation candidate à la

pertinence moyenne de la collection référentielle). Les travaux de Wanner et Alonso Ramos permettent d'espérer une modélisation similaire du sens de chaque fonction lexicale verbale. Avec un tel modèle, il sera alors possible de compléter l'encodage automatique des combinaisons extraites par Colex.

Nous avons volontairement simplifié la description de la stratégie d'encodage des combinaisons verbe + terme qui représente de nombreux défis pour la recherche en combinatoire lexicale. Les quelques pistes mentionnées ci-dessus démontrent cependant qu'elle pourrait être réalisée.

Index

D

E

F

I

L

M

O

P

R

S

T

U

V

Bibliographie

APRESYAN, YU. D., MEL'ČUK, I. A. et A. K. ŽOLKOVSKY (1969). « Semantics and Lexicography: Towards a New Type of Unilingual Dictionary » dans *Studies in Syntax and Semantics*, F. Kiefer (éditeur), Dordrecht/Boston, D. Reidel Publishing Company, pp. 1-33.

BEAUCHESNE, J. (2001). *Dictionnaire des cooccurrences*, Montréal, Guérin, 394 p.

BÉJOINT, H. et P. THOIRON (1992). « Macrostructure et microstructure dans un dictionnaire de collocations en langue de spécialité » dans *Terminologie et Traduction*, vol. 2-3, pp. 513-522.

BENSON, M. (1985). « Lexical Combinability » dans *Papers in Linguistics*, vol. 18, nº 1, Special issue: Advances in Lexicography, W. Frawley et R. Steiner (éditeurs), Boreal, Edmonton, Alberta, pp. 3-15.

BENSON, M. (1989). « The Structure of the Collocational Dictionary » dans *International Journal of Lexicography*, vol. 2, nº 1, pp. 1-13.

BENSON, M., BENSON, E. et R. ILSON (1986). *Lexicographic Description of English*, Amsterdam/Phildelphia, John Benjamins Publishing Company, 288 p.

BENSON, M., BENSON, E. et R. ILSON (1997). *The BBI Dictionary of English Word Combinations*, Amsterdam/Phildelphia, John Benjamins Publishing Company, 386 p.

BERRY-ROGGHE, G. (1973). « The computation of collocations and their relevance in lexical studies » dans *The Computer and Literary Studies*, A. J. Aitken *et al.* (éditeurs), Edinburgh, Edinburgh University Press, pp. 103-112.

BINON, J., VERLINDE, S., VAN DYCK, J. et A. BERTELS (2000). *Dictionnaire d'apprentissage du français des affaires*, Paris, Didier, 720 p.

CAIGNON, P. (2000). *Essential Lexicon in Accounting*, Saint-Laurent (Québec), Fides, 197 p.

CHANIER, T., FOUQUERÉ, C. et F. ISSAC (1995). « AlexiA : Un environnement d'aide à l'apprentissage lexical du français langue seconde » dans *Conférence Environnements Interactifs d'Apprentissage avec Ordinateur (EIAO'95)*, Eyrolles, Paris, pp. 79-90.

CHOUEKA, Y., KLEIN, S.T. et E. NEUWITZ (1983). « Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus » dans *Journal of the Association for Literary and Linguistic Computing*, vol. 4, pp. 34-38.

CHURCH, K., GALE, W., HANKS, P. et D. HINDLE (1991). Using Statistics in Lexical Analysis. Voir Zernik (1991), pp. 115-164.

CHURCH, K. W. et P. HANKS (1989). « Word Association Norms, Mutual Information, and Lexicography » dans *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 26-29 June 1989, Vancouver, Canada, pp. 76-83.

COHEN, B. (1986). *Lexique de cooccurrents Bourse Conjoncture économique*, Montréal, Linguatech, 125 p.

COHEN, B. (1992). « Méthodes de repérage et de classement des cooccurrents lexicaux » dans *Terminologie et traduction*, vol. 2-3, pp. 505-511.

Collins COBUILD Advanced Learner's English Dictionary, HarperCollins Publishers, Third Edition, 2001.

COWIE, A. P. (éditeur) (1998). *Phraseology: Theory, Analysis and Applications*, Oxford, Oxford University Press.

DAILLE, B. (1994). *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*, thèse de doctorat, Paris, Université de Paris 7, 228 p.

DAILLE, B. (2001). *Extraction de collocations à partir de textes*, tutoriel TALN 2001, 2-5 juillet 2001, Tours, 6 p.

DANCETTE, J. et M.-C. L'HOMME (2002). « The Gate to Knowledge in a Multilingual Specialized Dictionary: Using Lexical Functions for Taxonomic and Partitive Relations » dans *Proceedings, EURALEX 2002*, August 13-17 2002, Copenhagen (Denmark), pp. 597-606.

DANCETTE, J. et C. RÉTHORÉ (2000). *Dictionnaire analytique de la distribution; Analytical Dictionary of Retailing*, Montréal, Presses de l'Université de Montréal.

Dictionnaire d'Internet, de l'informatique et des télécommunications : anglais-français, Les publications du Québec, 2001.

FABRE, C. et C. FRÉROT (2002). « Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus » dans *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2002)*, Nancy, Tome 1, pp. 215-224.

FIRTH, J. R. (1957). « Modes of Meaning » dans *Papers in linguistics 1934-1951*, London, Oxford University Press, pp. 190-215.

FONTENELLE, T. (1997). *Turning a Bilingual Dictionary into a Lexical-Semantic Database*, Tübingen, Max Niemeyer Verlag, 328 p.

FRAWLEY, W. (1988). « New Forms of Specialized Dictionaries » dans *International Journal of Lexicography*, vol. 1, nº 3, pp. 189-213.

FREIBOTT, G. et U. HEID (1990). « Terminological and Lexical Knowledge for Computer-Aided Translation and Technical Writing » dans *TKE'90:*

Terminology and Knowledge Engineering, H. Czap et W. Nedobity (éditeurs), Frankfurt/M., Indeks Verlag, pp. 522-535.

GINGUAY, M. (2001). *Dictionnaire d'informatique anglais-français*, Dunod, Paris, 298 p.

GOLDMAN, J.P., NERIMA, L. et E. WEHRLI (2001). « Collocation Extraction Using a Syntactic Parser » dans *Proceedings of the Workshop on Collocations: Computational Extraction, Analysis and Exploitation, ACL-EACL 2001*, Toulouse, pp. 61-66.

GREFENSTETTE, G. et S. TEUFEL (1995). « Corpus-based Method for Automatic Identification of Support Verbs for Nominalizations » dans *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, March 27-31, 1995, Dublin, Ireland, pp. 98-103.

GROSS, M. (1982). « Une classification des phrases « figées » du français » dans *Revue québécoise de linguistique*, vol 11, nº 2, pp. 151-185.

GROSSMANN, F. et A. TUTIN (éditeurs) (2003). *Les collocations. analyse et traitement*, Publications Linguistiques, collection Travaux et Recherches en Linguistique Appliquée, Éditions De Werelt, Amsterdam, 142 p.

HALLIDAY, M. A. K. (1966). « Lexis as a Linguistic Level » dans *In Memory of J. R. Firth*, C. E. Bazell *et al.* (éditeurs), London, Longmans, pp. 148-162.

HALLIDAY, M. A. K. (2002). *On Grammar, Volume 1 in the Collected Works of M. A. K. Halliday*, J. Webster (éditeur), London and New York, Continuum, 442 p.

HAUSSMANN, F. J. (1979). « Un dictionnaire des collocations est-il possible ? » dans *Travaux de linguistique et de littérature*, vol. 17, nº 1, pp. 187-195.

HEID, U. (1992). « Décrire les collocations » dans *Terminologie et Traduction*, vol. 2-3, pp. 523-548.

HEID, U. (1996). Using Lexical Functions for the Extraction of Collocations from Dictionaries and Corpora. See Wanner (1996), pp. 115-146.

HEID, U. et G. FREIBOTT (1991). « Collocations dans une base de données terminologique et lexicale » dans *Meta*, vol 36, nº 1, pp. 77-91.

HORNBY, A. S. (2000). *Oxford Advanced Learner's Dictionary of Current English*, S. Wehmeier (éditeur), Oxford University Press, Sixth Edition, 1539 p.

KAHANE, S. et A. POLGUÈRE (2001). "Formal foundation of lexical functions" dans *Proceedings of the Workshop on Collocations: Computational Extraction, Analysis and Exploitation, ACL-EACL 2001*, Toulouse, pp. 8-15.

KILGARRIFF, A. (1996). "Which words are particularly characteristic of a text? A survey of statistical approaches" dans *Proceedings of the AISB Workshop on Language Engineering for Document Analysis and Recognition*, Sussex, pp. 33-40.

KILGARRIFF, A. et D. TUGWELL (2001). "WORD SKETCH: Extraction, Combination and Display of Significant Collocations for Lexicography" dans *Proceedings of the Workshop on Collocations: Computational Extraction, Analysis and Exploitation, ACL-EACL 2001*, Toulouse, pp. 32-38.

KRENN, B. et S. EVERT (2001). « Can we do better than frequency? A case study on extracting PP-verb Collocations » dans *Proceedings of the*

Workshop on Collocations: Computational Extraction, Analysis and Exploitation, ACL-EACL 2001, Toulouse, pp. 39-46.

LACROIX, U. (1958). *Les mots et les idées. Dictionnaire des termes cadrant avec les idées*, Paris, Fernand Nathan. 317 p.

LAINÉ, C. (1993). *Vocabulaire combinatoire de la CFAO mécanique*, Bureau de la traduction, Travaux publics et Services gouvernementaux Canada.

LA PORTE, M. de (1571). *Les Epithètes de M. de La Porte*, Paris, G. Buon.

Le Jargon Français. http://www.linux-france.org/prj/jargonf/

L'HOMME, M.-C. (2002). « Fonctions lexicales pour représenter les relations sémantiques entre termes » dans *T.A.L*, vol. 43, n° 1, pp. 19-41.

L'HOMME, M.-C. (2004a, à paraître). "Using Explanatory and Combinatorial Terminology to Describe Terms" dans *Selected Lexical and Grammatical Topics in the Meaning-Text Theory. In Honour of Igor Mel'cuk*, L. Wanner (éditeur), Amsterdam/Philadelphia, John Benjamins Publishing Company.

L'HOMME, M.-C. (2004b). « Sélection de termes dans un dictionnaire d'informatique : comparaison de corpus et critères lexico-sémantiques » dans *Proceedings of the Eleventh EURALEX International Congress*, July 6-10, 2004, Lorient, France, pp. 583-593.

L'HOMME, M.-C. et J. DANCETTE (2001). « Modélisation des relations sémantiques dans un dictionnaire spécialisé bilingue » dans *L'éloge de la différence : la voix de l'autre. Actes des 6e Journées du Réseau LTT*, A. Clas *et al.* (éditeurs), collection actualité scientifique, AUPELF-UREF, Paris, pp. 385-400.

LIN, D. (1998). "Extracting Collocations from Text Corpora" dans *First Workshop on Computational Terminology, COLING-ACL '98*, Montréal, pp. 57-63.

Longman Dictionary of Contemporary English, Addison-Wesley Pub Co, Third Edition, 1996.

LUDEWIG, P. (2001). « LogoTax – un outil exploratoire pour l'étude de collocations en corpus » dans *T.A.L*, vol. 42, n° 2, pp. 623-642.

MANNING, C. D. et H. SCHÜTZE (1999). *Foundations of Statistical Natural Language Processing*, Cambridge, Massachusetts, MIT Press, 680 p.

MARSHMAN, E. (2003). *Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie.* http://www.ling.umontreal.ca/lhomme/terminotique.html

MEL'ČUK, I. A. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. Voir Wanner (1996), pp. 37-102.

MEL'ČUK, I. A. (1998). Collocations and Lexical Functions. Voir Cowie (1998), pp. 23-53.

MEL'ČUK, I. A. (2003). « Collocations dans le dictionnaire » dans *Les écarts culturels dans les dictionnaires bilingues*, T. Szende (sous la dir. de), Paris, Honoré Champion, pp. 19-64.

MEL'ČUK, I.A., ARBATCHEWSKY-JUMARIE, N., ELNITSKY, L., IORDANSKAJA, L. et A. LESSARD (1984). *Dictionnaire explicatif et combinatoire du français contemporain – Recherches lexico-sémantiques I*, Montréal, Presses de l'Université de Montréal, 172 p.

MEL'ČUK, I. A., ARBATCHEWSKY-JUMARIE, N., DAGENAIS, L., ELNITSKY, L., IORDANSKAJA, L., LEFEBVRE, M.-N. et S. MANTHA (1988). *Dictionnaire explicatif et combinatoire du français contemporain – Recherches*

lexico-sémantiques II, Montréal, Presses de l'Université de Montréal, 332 p.

MEL'ČUK, I. A., ARBATCHEWSKY-JUMARIE, N., IORDANSKAJA, L. et S. MANTHA (1992). *Dictionnaire explicatif et combinatoire du français contemporain – Recherches lexico-sémantiques III*, Montréal, Presses de l'Université de Montréal, 323 p.

MEL'ČUK, I. A., ARBATCHEWSKY-JUMARIE, N., IORDANSKAJA, L., MANTHA, S. et A. POLGUÈRE (1999). *Dictionnaire explicatif et combinatoire du français contemporain – Recherches lexico-sémantiques IV*, Montréal, Presses de l'Université de Montréal, 347 p.

MEL'ČUK, I. A., CLAS, A. et A. POLGUÈRE (1995). *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve, Éditions Duculot/AUPELF-UREF, 256 p.

MEYER, I. et K. MACKINTOSH (1996). « The Corpus from a Terminographer's Viewpoint » dans *International Journal of Corpus Linguistics*, vol. 1, n° 2, pp. 257-285.

MEYNARD, I. (1997). *Méthode de consignation dans un outil HTML des combinaisons lexicales spécialisées : étude basée sur 15 termes français et 15 termes anglais tirés du domaine de l'Internet*, mémoire de maîtrise, Montréal, Université de Montréal.

MEYNARD, I. (2000). *Internet : répertoire bilingue de combinaisons lexicales spécialisées*, Brossard (Québec), Linguatech, 207 p.

PANTEL, P. et D. LIN (2000). « Word-for-Word Glossing with Contextually Similar Words » dans *Proceedings of ANLP-NAACL 2000*, Seattle, Washington, pp. 78-85.

PAVEL, S. et M. BOILEAU (2003). *Vocabulaire combinatoire de l'imagerie fractale*, Bureau de la traduction, Travaux publics et Services gouvernementaux Canada.

PEARSON, J. (1998). *Terms in Context*, Amsterdam/Philadelphia, John Benjamins Publishing Company.

POLGUÈRE, A. (2000). « Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French » dans *Proceedings of EURALEX'2000*, Stuttgart, pp. 517-527.

POLGUÈRE, A. (2003a). *Lexicologie et sémantique lexicale. Notions fondamentales*, collection « Paramètres », Montréal, Les Presses de l'Université de Montréal, 260 p.

POLGUÈRE, A. (2003b). Collocations et fonctions lexicales : pour un modèle d'apprentissage. Voir Grossmann et Tutin (2003), pp. 117-133.

POLGUÈRE, A. (2003c). « Étiquetage sémantique des lexies dans la base de données DiCo » dans *T.A.L*, vol. 44, n° 2, pp. 39-68.

RAT, M. (1999). *Dictionnaire des expressions et locutions traditionnelles*, Larousse, Paris, 448 p.

REY, A. et S. CHANTREAU (1999). *Dictionnaire des expressions et locutions*, Dictionnaires Le Robert, collection « les usuels », Paris, 888 p.

ROBERT, P. (1984). *Petit Robert 1 : Dictionnaire alphabétique et analogique de la langue française*, rédaction dirigée par A. Rey et J. Rey-Debove, Dictionnaires Le Robert, Paris, 2171 p.

ROUAIX, P. (1989). *Trouver le mot juste : dictionnaire des idées suggérées par les mots*, Le Livre de Poche, n° 7939, 538 p.

SCHMID, P. et C. GDANIEC (1996). "Evolution of the Logos Grammar: System Design and Development Methodology" dans *Proceedings of the*

Second Conference of the Association for Machine Translation in the Americas, 2-5 October 1996, Montreal, Canada, pp. 86-95.

SCOTT, B. E. (1999). « Linguistic and Computational Motivations for the LOGOS Machine Translation System: An Overview » dans *Hybrid Approaches to Machine Translation*, Streiter *et al.* (éditeurs), IAI Working Paper No. 36, Institute of Applied Information Sciences, Saarbrücken, Germany. http://www.iai.uni-sb.de/iaien/iaiwp/index.htm

SINCLAIR, J. (1987). « Collocation: a progress report » dans *Language Topics: Essays in honour of Michael Halliday*, R. Steele et T. Threadgold (éditeurs), Amsterdam/Phildelphia, John Benjamins Publishing Company, pp. 319-331.

SINCLAIR, J., JONES, S. et R. DALEY (1970). *English Lexical Studies: Report to OSTI on Project C/LP/08*, Department of English, University of Birmingham.

SMADJA, F. (1993). « Retrieving Collocations from Text: Xtract » dans *Computational Linguistics* , vol. 19, n° 1, pp. 143-177.

VERLINDE, S., SELVA, T. et J. BINON (2003). Les collocations dans les dictionnaires d'apprentissage : repérage, présentation et accès. Voir Grossmann et Tutin (2003), pp. 105-115.

WANNER, L. (éditeur) (1996). *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/Philadelphia, John Benjamins Publishing Company, 355 p.

WANNER, L. et M. ALONSO RAMOS (2001). *Vers une approche sémantique pour l'identification des collocations en corpus*, communication aux Journées d'Étude de l'ATALA, Samedi 13 janvier 2001, Paris, 5 p.

ZERNIK, U. (éditeur) (1991). *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Hillsdale, New Jersey, Lawrence Erlbaum Associates, 429 p.

Annexes

Annexe A – Documents composant le corpus

Nom	*Sous-domaine*	*Niveau de spéc.*	*Type de document*	*Nbre de mots*
32bits	Comprendre l'inform.	Technique	Revue spécialisée	88893
anniversary	Comprendre l'inform.	Vulgarisation	Revue spécialisée	1307
ataleo	Progr. et réseaux	Très technique	Autre ouvrage	2952
buildingpc	Matériel	Didactique	Article de vulgarisation	23095
cache	Matériel	Didactique	Article de vulgarisation	2380
CGI1	Progr. et réseaux	Didactique	Manuel de cours	10476
CGI2	Progr. et réseaux	Didactique	Manuel de cours	9073
CGI3	Progr. et réseaux	Didactique	Manuel de cours	9112
CGI6	Progr. et réseaux	Didactique	Manuel de cours	7774
chess	Comprendre l'inform.	Didactique	Manuel technique	2870
china	Comprendre l'inform.	Vulgarisation	Revue spécialisée	2983
crash	Comprendre l'inform.	Vulgarisation	Article de vulgarisation	1830
cremag1	Progr. et réseaux	Technique	Revue spécialisée	3043
cremag2	Progr. et réseaux	Technique	Revue spécialisée	1161
cremag3	Système d'exploitation	Technique	Revue spécialisée	1433
cremag4	Logiciel	Technique	Revue spécialisée	1087
cremag5	Internet	Technique	Revue spécialisée	1729
cremag6	Internet	Technique	Revue spécialisée	1372
cremag7	Logiciel	Technique	Revue spécialisée	1274
cyber	Internet	Vulgarisation	Revue spécialisée	2668
datawa	Matériel	Technique	Autre ouvrage	4133
device01	Matériel	Didactique	Manuel de cours	1577
device02	Matériel	Didactique	Manuel de cours	940
device03	Matériel	Didactique	Manuel de cours	1093
email	Internet	Didactique	Article de vulgarisation	2410
familyp2	Internet	Vulgarisation	Revue spécialisée	37125
Firewall	Comprendre l'inform.	Vulgarisation	Autre ouvrage	705
generation	Progr. et réseaux	Technique	Article de vulgarisation	1388
getstarted	Comprendre l'inform.	Vulgarisation	Autre ouvrage	6878
guide	Comprendre l'inform.	Didactique	Article de vulgarisation	18961
hardware	Matériel	Didactique	Manuel de cours	15043
HP	Système d'exploitation	Technique	Manuel technique	1232

html	Internet	Didactique	Manuel de cours	8028
inputdevice	Matériel	Technique	Article de vulgarisation	2245
integr	Progr. et réseaux	Très technique	Autre ouvrage	6642
integration	Progr. et réseaux	Technique	Article de vulgarisation	2298
internet	Internet	Didactique	Article de vulgarisation	27195
introwin	Système d'exploitation	Didactique	Autre ouvrage	3643
java	Progr. et réseaux	Didactique	Article de vulgarisation	4987
keyboard	Matériel	Didactique	Article de vulgarisation	2015
maintaining	Comprendre l'inform.	Didactique	Article de vulgarisation	2114
mandel	Comprendre l'inform.	Vulgarisation	Manuel de cours	41146
McAfee	Logiciel	Technique	Autre ouvrage	970
memory	Matériel	Didactique	Article de vulgarisation	1210
micropro	Matériel	Didactique	Article de vulgarisation	2737
motherboard	Matériel	Didactique	Article de vulgarisation	2818
mydocuments	Système d'exploitation	Vulgarisation	Revue spécialisée	1606
opersystem	Système d'exploitation	Didactique	Article de vulgarisation	1017
overclocking	Matériel	Didactique	Article de vulgarisation	1577
pc101	Comprendre l'inform.	Didactique	Article de vulgarisation	4847
pcmag2	Matériel	Technique	Revue spécialisée	58699
pcwork	Comprendre l'inform.	Didactique	Article de vulgarisation	2972
progC	Progr. et réseaux	Didactique	Article de vulgarisation	18433
raytracer	Logiciel	Technique	Article de vulgarisation	2400
scanner	Matériel	Vulgarisation	Revue spécialisée	3793
searchnet	Internet	Technique	Autre ouvrage	12338
security01	Progr. et réseaux	Technique	Revue spécialisée	854
security02	Comprendre l'inform.	Technique	Autre ouvrage	1335
soundcard	Matériel	Didactique	Article de vulgarisation	1216
speaker	Matériel	Didactique	Article de vulgarisation	1089
storage	Comprendre l'inform.	Vulgarisation	Revue spécialisée	1689
technology	Comprendre l'inform.	Vulgarisation	Autre ouvrage	5414
tuneup	Comprendre l'inform.	Didactique	Autre ouvrage	1659
virusinfo01	Comprendre l'inform.	Vulgarisation	Autre ouvrage	2553
virusinfo02	Comprendre l'inform.	Vulgarisation	Autre ouvrage	5250
virusinfo03	Comprendre l'inform.	Vulgarisation	Autre ouvrage	3464
visual	Matériel	Vulgarisation	Manuel de cours	13770
voice	Logiciel	Technique	Revue spécialisée	2602
w2kinst	Système d'exploitation	Didactique	Article de vulgarisation	3474
web	Internet	Vulgarisation	Manuel de cours	996
web2	Internet	Didactique	Article de vulgarisation	4992

webDev1	Internet	Didactique	Manuel de cours	4976
webDev2	Internet	Didactique	Manuel de cours	5217
webDev3	Internet	Didactique	Manuel de cours	2794
webDev4	Internet	Didactique	Manuel de cours	3280
white1	Matériel	Vulgarisation	Manuel de cours	19325
winmag1	Comprendre l'inform.	Vulgarisation	Revue spécialisée	255
winmag10	Matériel	Technique	Revue spécialisée	1429
winmag2	Comprendre l'inform.	Vulgarisation	Revue spécialisée	270
winmag3	Matériel	Technique	Revue spécialisée	743
winmag4	Comprendre l'inform.	Vulgarisation	Revue spécialisée	4663
winmag6b	Comprendre l'inform.	Vulgarisation	Revue spécialisée	694
winmag6b	Internet	Vulgarisation	Revue spécialisée	1239
winmag7	Comprendre l'inform.	Vulgarisation	Revue spécialisée	697
winmag8	Comprendre l'inform.	Vulgarisation	Revue spécialisée	745
winmag9	Matériel	Technique	Revue spécialisée	1477
winXP	Système d'exploitation	Vulgarisation	Article de journal	1623
wired1	Comprendre l'inform.	Vulgarisation	Revue spécialisée	11731
wired2	Comprendre l'inform.	Vulgarisation	Revue spécialisée	1731
zabus	Matériel	Technique	Manuel de cours	2859
zacpumem	Matériel	Technique	Manuel de cours	5090
zanorton	Matériel	Technique	Manuel de cours	2683
zaplatyp	Matériel	Technique	Manuel de cours	3428

Annexe B - Règles de la grammaire Obj1

```
"<>1 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$10, $8]); ',

"<>2 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass) (($adv)*)($n_agent)
(($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$10, $8]); ',

"<>3 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($vi_act) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVI([$10, $15]); ',

"<>4 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($vt_act) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$10, $15]); ',

"<>5 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prev_act) ($aux_pass_compl)
($v_pass_compl) (($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$20, $8]); ',

"<>6 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prev_act) ($aux_pass_compl)
($v_pass_compl) (($adv)*)($n_agent) (($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$20, $8]); ',

"<>7 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prev_act)
($prep)?($vi_compl) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVI([$20, $25]); ',

"<>8 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prev_act)
($prep)?($vt_compl) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$20, $25]); ',

"<>9 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prevDITR_act) ($n_dobj)
($aux_pass_compl) ($v_pass_compl) (($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$18, $15]); ',

"<>10 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prevDITR_act) ($n_dobj)
($aux_pass_compl) ($v_pass_compl) (($adv)*)($n_agent)
(($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$18, $15]); ',

"<>11 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prevDITR_act) ($anyNoun)
($prep)?($vi_compl) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVI([$17, $22]); ',
```

```
"<>12 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prevDITR_act) ($anyNoun)
($prep)?($vt_compl) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$17, $22]); ',

"<>13 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prevDITR_pass)
($prep)?($vi_compl) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$16, $21]); ',

"<>14 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prevDITR_pass)
($prep)?($vt_compl) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$16, $21]); ',

"<>15 ($BOP) ($vi_imper) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$6, $11]); ',

"<>16 ($BOP) ($vt_imper) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$6, $11]); ',

"<>17 ($BOP) ($v_prev_imper) ($prep)?($vi_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$16, $21]); ',

"<>18 ($BOP) ($v_prev_imper) ($prep)?($vt_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$16, $21]); ',

"<>19 ($BOP) ($v_prevDITR_imper) ($anyNoun) ($prep)?($vi_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$13, $18]); ',

"<>20 ($BOP) ($v_prevDITR_imper) ($anyNoun) ($prep)?($vt_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$13, $18]); ',

"<>21 ($BOP) ($vi_ing) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$8, $12]); ',

"<>22 ($BOP) ($vt_ing) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$8, $12]); ',

"<>23 ($BOP) ($v_prev_ing) ($prep)?($vi_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$14, $19]); ',

"<>24 ($BOP) ($v_prev_ing) ($prep)?($vt_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$14, $19]); ',

"<>25 ($BOP) ($v_prevDITR_ing) ($anyNoun) ($prep)?($vi_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
```

```
  =>  ' &incrFreqVI([$13, $18]); ',

"<>26 ($BOP) ($v_prevDITR_ing) ($anyNoun) ($prep)?($vt_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$13, $18]); ',

"<>27 ($BOP) ($n_prev) ($vi_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$8, $12]); ',

"<>28 ($BOP) ($n_prev) ($vt_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$8, $12]); ',

"<>29 ($BOP) ($n_prev) ($v_prev_compl) ($prep)?($vi_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$14, $19]); ',

"<>30 ($BOP) ($n_prev) ($v_prev_compl) ($prep)?($vt_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$14, $19]); ',

"<>31 ($BOP) ($n_prev) ($v_prevDITR_compl) ($anyNoun) ($prep)?($vi_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$13, $18]); ',

"<>32 ($BOP) ($n_prev) ($v_prevDITR_compl) ($anyNoun) ($prep)?($vt_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$13, $18]); ',

"<>33 ($BOP) ($vi_adv) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$6, $11]); ',

"<>34 ($BOP) ($vt_adv) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$6, $11]); ',

"<>35 ($BOP) ($v_prev_adv) ($prep)?($vi_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$16, $21]); ',

"<>36 ($BOP) ($v_prev_adv) ($prep)?($vt_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$16, $21]); ',

"<>37 ($BOP) ($v_prevDITR_adv) ($anyNoun) ($prep)?($vi_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$13, $18]); ',

"<>38 ($BOP) ($v_prevDITR_adv) ($anyNoun) ($prep)?($vt_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
```

```
  =>  ' &incrFreqVT([$13, $18]); ',

"<>39 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act) (($n_obj)*)($vi_adv)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$15, $20]); ',

"<>40 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act) (($n_obj)*)($vt_adv)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$15, $20]); ',

"<>41 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass) (($n_obj)*)($vi_adv)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$15, $20]); ',

"<>42 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass) (($n_obj)*)($vt_adv)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$15, $20]); ',

"<>43 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act)
(($n_obj)*)($v_prev_adv) ($prep)?($vi_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$25, $30]); ',

"<>44 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act)
(($n_obj)*)($v_prev_adv) ($prep)?($vt_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$25, $30]); ',

"<>45 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act)
(($n_obj)*)($v_prevDITR_adv) ($anyNoun) ($prep)?($vi_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$22, $27]); ',

"<>46 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act)
(($n_obj)*)($v_prevDITR_adv) ($anyNoun) ($prep)?($vt_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$22, $27]); ',

"<>47 ($BOP) ($v_imper) (($n_obj)*)($vi_adv) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$11, $16]); ',

"<>48 ($BOP) ($v_imper) (($n_obj)*)($vt_adv) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$11, $16]); ',

"<>49 ($BOP) ($v_imper) (($n_obj)*)($v_prev_adv) ($prep)?($vi_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$21, $26]); ',
```

```
"<>50 ($BOP) ($v_imper) (($n_obj)*)($v_prev_adv) ($prep)?($vt_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$21, $26]); ',

"<>51 ($BOP) ($v_imper) (($n_obj)*)($v_prevDITR_adv) ($anyNoun)
($prep)?($vi_compl) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVI([$18, $23]); ',

"<>52 ($BOP) ($v_imper) (($n_obj)*)($v_prevDITR_adv) ($anyNoun)
($prep)?($vt_compl) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$18, $23]); ',

"<>53 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass)
(($adv)*)(($n_obj)*)($conj) ($anyAux)?($v_pass)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$21, $8]); ',

"<>54 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass) (($adv)*)($n_iobj)
(($n_iobj)*)($conj) ($anyAux)?($v_pass)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$10, $8]); ',

"<>55 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass)
(($adv)*)(($n_obj)*)($conj) ($anyAux)?($v_pass) (($adv)*)($n_agent)
(($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$21, $8]); ',

"<>56 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass) (($adv)*)($n_agent)
(($n_iobj)*)($conj) ($anyAux)?($v_pass)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$10, $8]); ',

"<>57 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act)
(($adv)*)(($n_obj)*)($conj) ($anyAux)?($vi_act) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVI([$21, $26]); ',

"<>58 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act)
(($adv)*)(($n_obj)*)($conj) ($anyAux)?($vt_act) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  => ' &incrFreqVT([$21, $26]); ',

"<>59 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($vi_act) (($adv)*)($n_obj)
(($n_iobj)*)($conj) ($anyAux)?($v_act)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  => ' &incrFreqVI([$10, $15]); ',

"<>60 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($vt_act) (($adv)*)($n_obj)
(($n_iobj)*)($conj) ($anyAux)?($v_act)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
```

```
  =>  ' &incrFreqVT([$10, $15]); ',

"<>61 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prev_act)
($prep)?($v_compl) (($adv)*)(($n_obj)*)($conj) ($prep)?($vi_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$31, $36]); ',

"<>62 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prev_act)
($prep)?($v_compl) (($adv)*)(($n_obj)*)($conj) ($prep)?($vt_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$31, $36]); ',

"<>63 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prev_act)
($prep)?($vi_compl) (($adv)*)($n_obj) (($n_iobj)*)($conj)
($prep)?($v_compl) (($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$20, $25]); ',

"<>64 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prev_act)
($prep)?($vt_compl) (($adv)*)($n_obj) (($n_iobj)*)($conj)
($prep)?($v_compl) (($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$20, $25]); ',

"<>65 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prevDITR_act) ($anyNoun)
($prep)?($v_compl) (($adv)*)(($n_obj)*)($conj) ($prep)?($vi_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$28, $33]); ',

"<>66 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prevDITR_act) ($anyNoun)
($prep)?($v_compl) (($adv)*)(($n_obj)*)($conj) ($prep)?($vt_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$28, $33]); ',

"<>67 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prevDITR_act) ($anyNoun)
($prep)?($vi_compl) (($adv)*)($n_obj) (($n_iobj)*)($conj)
($prep)?($v_compl) (($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$17, $22]); ',

"<>68 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prevDITR_act) ($anyNoun)
($prep)?($vt_compl) (($adv)*)($n_obj) (($n_iobj)*)($conj)
($prep)?($v_compl) (($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$17, $22]); ',

"<>69 ($BOP) ($v_imper) (($adv)*)(($n_obj)*)($conj) ($vi_imper)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$16, $21]); ',

"<>70 ($BOP) ($v_imper) (($adv)*)(($n_obj)*)($conj) ($vt_imper)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$16, $21]); ',
```

```
"<>71 ($BOP) ($vi_imper) (($adv)*)($n_obj) (($n_iobj)*)($conj) ($v_imper)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$6, $11]); ',

"<>72 ($BOP) ($vt_imper) (($adv)*)($n_obj) (($n_iobj)*)($conj) ($v_imper)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$6, $11]); ',

"<>73 ($BOP) ($v_ing) (($adv)*)(($n_obj)*)($conj) ($vi_ing)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$17, $21]); ',

"<>74 ($BOP) ($v_ing) (($adv)*)(($n_obj)*)($conj) ($vt_ing)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$17, $21]); ',

"<>75 ($BOP) ($vi_ing) (($adv)*)($n_obj) (($n_iobj)*)($conj) ($v_ing)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$8, $12]); ',

"<>76 ($BOP) ($vt_ing) (($adv)*)($n_obj) (($n_iobj)*)($conj) ($v_ing)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$8, $12]); ',

"<>77 ($BOP) ($n_prev) ($v_compl) (($adv)*)(($n_obj)*)($conj) ($vi_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$17, $21]); ',

"<>78 ($BOP) ($n_prev) ($v_compl) (($adv)*)(($n_obj)*)($conj) ($vt_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$17, $21]); ',

"<>79 ($BOP) ($n_prev) ($vi_compl) (($adv)*)($n_obj) (($n_iobj)*)($conj)
($v_compl) (($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$8, $12]); ',

"<>80 ($BOP) ($n_prev) ($vt_compl) (($adv)*)($n_obj) (($n_iobj)*)($conj)
($v_compl) (($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$8, $12]); ',

"<>81 ($BOP) ($v_compl) (($adv)*)(($n_obj)*)($conj) ($vi_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$15, $19]); ',

"<>82 ($BOP) ($v_compl) (($adv)*)(($n_obj)*)($conj) ($vt_compl)
(($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$15, $19]); ',

"<>83 ($BOP) ($vi_compl) (($adv)*)($n_obj) (($n_iobj)*)($conj) ($v_compl)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
```

```
  =>  ' &incrFreqVI([$6, $10]); ',

"<>84 ($BOP) ($vt_compl) (($adv)*)($n_obj) (($n_iobj)*)($conj) ($v_compl)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$6, $10]); ',

"<>85 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act) (($n_obj)*)($v_adv)
(($adv)*)(($n_obj)*)($conj) ($vi_adv) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$25, $30]); ',

"<>86 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act) (($n_obj)*)($v_adv)
(($adv)*)(($n_obj)*)($conj) ($vt_adv) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$25, $30]); ',

"<>87 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act) (($n_obj)*)($vi_adv)
(($adv)*)($n_obj) (($n_iobj)*)($conj) ($v_adv)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$15, $20]); ',

"<>88 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act) (($n_obj)*)($vt_adv)
(($adv)*)($n_obj) (($n_iobj)*)($conj) ($v_adv)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$15, $20]); ',

"<>89 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass) (($n_obj)*)($v_adv)
(($adv)*)(($n_obj)*)($conj) ($vi_adv) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$25, $30]); ',

"<>90 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass) (($n_obj)*)($v_adv)
(($adv)*)(($n_obj)*)($conj) ($vt_adv) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$25, $30]); ',

"<>91 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass) (($n_obj)*)($vi_adv)
(($adv)*)($n_obj) (($n_iobj)*)($conj) ($v_adv)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$15, $20]); ',

"<>92 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass) (($n_obj)*)($vt_adv)
(($adv)*)($n_obj) (($n_iobj)*)($conj) ($v_adv)
(($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$15, $20]); ',

"<>93 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass) ($conj)
($anyAux)?($v_pass) (($adv)*)(($n_obj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$10, $8]); ',
```

```
"<>94 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass) ($conj)
($anyAux)?($v_pass) (($adv)*)($n_agent) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$10, $8]); ',

"<>95 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($vi_act) ($conj)
($anyAux)?($v_act) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$10, $21]); ',

"<>96 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($vt_act) ($conj)
($anyAux)?($v_act) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$10, $21]); ',

"<>97 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prev_act)
($prep)?($vi_compl) ($conj) ($prep)?($v_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$20, $31]); ',

"<>98 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prev_act)
($prep)?($vt_compl) ($conj) ($prep)?($v_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$20, $31]); ',

"<>99 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prevDITR_act) ($anyNoun)
($prep)?($vi_compl) ($conj) ($prep)?($v_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$17, $28]); ',

"<>100 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_prevDITR_act) ($anyNoun)
($prep)?($vt_compl) ($conj) ($prep)?($v_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$17, $28]); ',

"<>101 ($BOP) ($vi_imper) ($conj) ($v_imper) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$6, $16]); ',

"<>102 ($BOP) ($vt_imper) ($conj) ($v_imper) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$6, $16]); ',

"<>103 ($BOP) ($vi_ing) ($conj) ($v_ing) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$8, $16]); ',

"<>104 ($BOP) ($vt_ing) ($conj) ($v_ing) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$8, $16]); ',

"<>105 ($BOP) ($n_prev) ($vi_compl) ($conj) ($v_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
```

```
  =>  ' &incrFreqVI([$8, $16]); ',

"<>106 ($BOP) ($n_prev) ($vt_compl) ($conj) ($v_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$8, $16]); ',

"<>107 ($BOP) ($vi_compl) ($conj) ($v_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$6, $14]); ',

"<>108 ($BOP) ($vt_compl) ($conj) ($v_compl) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$6, $14]); ',

"<>109 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act) (($n_obj)*)($vi_adv)
($conj) ($v_adv) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$15, $25]); ',

"<>110 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_act) (($n_obj)*)($vt_adv)
($conj) ($v_adv) (($adv)*)($n_obj) (($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$15, $25]); ',

"<>111 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass)
(($n_obj)*)($vi_adv) ($conj) ($v_adv) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVI([$15, $25]); ',

"<>112 ($BOP) (($adv)*)($anyNoun) ($anyAux)?($v_pass)
(($n_obj)*)($vt_adv) ($conj) ($v_adv) (($adv)*)($n_obj)
(($n_iobj)*)(($adv)*)($EOP)"
  =>  ' &incrFreqVT([$15, $25]); '
```