UNIVERSITE KASDI MERBAH-OUARGLA

Faculté des Lettres et des langues

Département des Langues Etrangères

N°d'ordre :
N° de série :

Ecole Doctorale de Français

Antenne de l'Université Kasdi Merbah-Ouargla

Mémoire

Pour l'obtention du diplôme de

MAGISTER DE FRANÇAIS

***Option :* Sciences du langage**

Par : TAIBAOUI Mohammed

Thème :

Enjeux linguistiques des expressions figées dans les textes journalistiques : Pour une approche automatique (TAL).

Directeur de recherche :

Dr. Salah KHENNOUR

Soutenu publiquement le 07/07/2009

Devant le jury composé de :

-	**Dr. Bachir BENSALAH**	**Président**	**Université de Biskra**
-	**Dr. Salah KHENNOUR**	**Rapporteur**	**Université de Ouargla**
-	**Dr. Samir ABDELHAMID**	**Examinateur**	**Université de Batna**
-	**Dr. Tarek BENZEROUAL**	**Examinateur**	**Université de Batna**

Remerciements

Je souhaite remercier très vivement, toutes les personnes qui ont pu m'aider et qui sont intervenues d'une manière ou d'une autre dans la réalisation de ce mémoire.

Je suis très reconnaissant, en particulier en vers mon enseignant et directeur de recherche le docteur Khennour Salah, pour sa disponibilité, son écoute et son aide.

Je tiens à remercier, également tous mes enseignants, notamment le professeur Foudil Dahou et le docteur Rachid Raissi qui m'a beaucoup soutenu aux moments les plus délicats.

Un grand merci très chaleureux, aussi aux collègues et amis qui m'ont beaucoup aidé et encouragé, tout au long de ce travail.

Je dédie ce travail à l'âme de mon père et à ma chère mère à qui je dois tout.

Table des matières.

Chapitre III

Approches linguistiques et enjeux informatiques des expressions figées .

Chapitre IV

Traitement automatique des expressions figées

Mots-clés: *figement, expressions figées, mots composes, idiomatique, phraséologie, non-compositionalité, restrictions syntaxiques, texts journalistiques , traitement automatique des langues (TAL), LADL(laboratoire d'automatique documentaire et linguistique), application des ressources lexicales (dictionnaires électroniques, tables de lexique-grammaire).*

Résumé

Les expressions figées résultent du processus de l'intégration d'une séquence du discours dans le système de la langue. Ces suites constituent des nouvelles unités lexicales dont le sens est non-compositionnel. Cette étude montre, par ailleurs les différentes contraintes sémantiques et morphosyntaxiques que présentent ces syntagmes et met l'accent sur le fait de la mémorisation de ces séquences qui leur accordent un statut rituel et historique.

Cette recherche consiste à une approche au moyen des outils du traitement automatique des langues (TAL), dans le but de décrire les expressions figées dans un corpus de textes journalistique . Nous nous appuyons sur la théorie de lexique-grammaire et les études de Maurice Gross, notamment celle réalisé dans le cadre des recherches de LADL (laboratoire d'automatique documentaire et linguistique). Nous procédons, en effet à une méthode qui s'appuie sur l'application automatique des ressources lexicales constituées de dictionnaires électronique et de tables de lexique-grammaire, décrivant les locutions et les expressions figées. Pour ce faire, nous aurons recours aux applications *Intex* et *Nooj* , conçues par Max Silberztein afin de décrire les locutions et les mots composés dans notre corpus de textes journalistiques et comparer cette situations à celle dans un autre type de textes.

Cette étude se veut par conséquent, une confirmation de la pertinence du phénomène du figement comme étant une caractéristique inhérente aux langues naturelles et de l'intérêt des méthodes automatique pour la description et l'étude des expressions figées.

---

keywords : *fixedness(freezing), frozen expressions, compound words, idiomatic, phraseology, non compositionality, syntactic restrictions, journalistic texts, natural language processing (NLP), LADL (Automatic Control Laboratory Documentary) , application of, lexical resources(electronic dictionaries, tables lexicon grammar).*

Abstract

Frozen expressions result from the process of integrating a sequence of speech in the system of language. These suites are new lexical units whose meaning is not compositional.this study shows,morevoer different morphosyntactic and semantic constraints of these phrases and emphasizes the fact of the memory of these sequences which give them a ritual and historic status.

Thise research is an approach using the tools of natural languge processing (NLP),to describe the fixed expressions in a cropus of journalistic textes.We rely on the theory of lexico-grammar and studies of Maurice Gross,including that done in the framework of reserch LADL(laboratoire d'automatoique doumentaire et linguistique).We proceed,in effect a method that relies on the automatic application of lexical ressources consist of electronic dictionaries and tables of lexicon-grammar,describing the phrases and expressions frozen. To do this,we will use applications of *Intex* and *Nooj* designed by Max sliberztein to describe the phrases and compound words in our corpus of journalistic texts and compare this situation with that in another type of texts.

This study is therefore a confirmation of the relevance of the phenomenon of the freeze (fixedness) as a characteristic inherent to natural languages and the interst of automatic methods for description and study of fixed expessions.

الكلمات المفتاحية. الجمود،العبارات الجامدة، الكلمات المركبة، إديوم، جملية، لا تركيبية، إمتناعات صرفية، نصوص صحفية، المعالجة الآلية للغات، (LADL) (مخبر آلية الوثائق واللسانيات)، تطبيق الموارد اللفضية (المعاجم الإلكترونية والجداول اللفظيةالنحوية).

ملخص

إنّ العبارات الجامدة هي نتاج إدماج قطعة من الخطاب في نظام التتابع من الكلمات يشكل وحدات لفظية جديدة ذات دلالة لا تركيبية . هاته الدراسة تبين مختلف الضوابط الدلالية والتركيبية- النحوية التي تطرحها هاته التراكيب وتنطرق لفعل الذاكرة التي تحفظ هاته القطع اللفظية مما يعطيها بعدا تقليديا وتاريخيا.

هذا البحث يتمثل في مقاربة بواسطة أدواة المعالجة الآلية للغوات الطبيعية(TAL) تهدف الى توصيف حضور العبارات الجامدة في حالة النصوص الصحفية. نعتمد هنا على نظرية الألفاظ النحوية (lexique grammaire) ودراسات مريس كروس،وبالأخص تلك المنجزة في إطار الأبحاث التي يقوم بها مخبر LADL بفرنسا (مخبر آلية الوثائق واللسانيات) . نسلك هنا منهجا يعتمد على التطبيق الآلي للموارد اللفظية المتمثلة في المعاجم الإلكترونية والجداول اللفظية – النحوية التي تصف المقولات والعبارات الجامدة من أجل هذا نقوم بالإستعانة ببرنامجين آليين هما إينتاكس *Intex* ونوج *Nooj* المصممين من طرف ماكس سلبرشتاين وذلك لأجل دراسة المقولات والكلمات المركبة في موضوع الدراسة المتمثل في نصوص صحفية ومقارنة هاته الحالة مع غيرها في أصناف أخرى من النصوص.

هاته الدراسة تكون بنتيجة تحقيق لأهمية حضور ظاهرة الجمود(figement) كخاصية أصلية للغات الطبيعية ولأهمية المناهج آلآلية لوصف ودراسة العبارات الجامدة.

Introduction générale

Chapitre I

Le figement dans la langue

Chapitre II

Les critères du figement

Chapitre III

Approches linguistiques et enjeux informatiques des expressions figées

Chapitre IV

Traitement automatique des expressions figées

Conclusion générale

Introduction générale

Choix du thème

Suivant la société, l'économie et la science, les langues naturelles connaissent depuis un siècle un développement sans précédant. En effet on avait besoin de dénommer de centaines de milliers d'objets et de concepts nouveaux et pour ce faire, on a eu recours à de différents moyens. On pourrait forger de nouvelles unités lexicales simples en combinant des lettres mais cette possibilité tend à être évitée puisque les langues connaissent une certaines stabilité à ce niveau *: « cette possibilité n'est guère exploitée par les langues, qui sont avant tout des réalités historiques »*[1]. En fait, nous sommes en présence de deux manières avec lesquelles se développe le trésor lexical des langues naturelles. Soit, on attribue des sens nouveaux aux mots déjà existants, ce qui favorise le caractère polysémique des unités lexicales et complique les messages; ou on fait recours à des combinaisons d'éléments lexicaux construits au moyen d'affixation (dérivation), ou bien, on crée des nouvelles dénominations en alliant des unités lexicales préexistantes (composition). Notre thème correspond à ce dernier procédé à savoir la composition et précisément aux expressions figées qui consistent à une réalisation du fait de la composition.

Nous étudions au cours de ce travail les expressions figées et le phénomène du figement qui font un objet primordial pour les linguistes qui travaillent dans le domaine de la phraséologie. Ces séquences constituent, d'ailleurs, un grand intérêt, aussi bien pour les apprenants que pour les simples usagers de la langue. En effet; si l'on accorde tout cet intérêt au phénomène du figement, c'est que ces suites de mots que nous appellerons tantôt des expressions figées, tantôt des mots composés ne représentent nullement un phénomène marginal de la langue. On estime que ces suites représentent entre 20et 30 % d'un texte donné et G. Gross fait remarquer qu'il y a *« près de 200 000 noms composés ou dont la combinatoire n'est pas libre, près de 15 000 adjectifs et au moins 30000 verbes figés »*[2]

Nous partons, dans cette étude de la constatation d'une abondance des termes utilisés pour décrire ces groupes de mots, parmi lesquels nous citons à titre d'exemple:

[1] GROOS, Gaston., *Les Expressions figées en français*, Ophrys, France, 1996, p. 161.
[2] GROOS, Gaston, « *Du bon usage de la notion de locution* », La locution entre langue et usages, ENS éditions, Fontenay Saint-Cloud.1997, p. 202

locutions, mots composés, idiotismes, clichés, proverbes, dictons, phrases figées, phrases toute faites, phraséologismes, phrasème etc. L'emploi de ces termes est, relativement équivoque mais il correspond, dans la plupart des cas au fait qu'une expression soit mémorisée par les locuteurs d'une langue et que son utilisation se veut conventionnelle et partagée de la plupart d'eux.

En fait, ce travail s'inscrit dans le cadre des études qu'on appelle généralement phraséologiques et porte en particulier sur le figement, lequel a été longtemps ignoré dans les études linguistiques, et qui commence à occuper une place de choix dans les préoccupations de la recherche actuelle, et ce en raison de son importance pour une meilleur connaissance des systèmes linguistiques. En effet Les études actuelles, pour récentes qu'elles soient, montrent qu'il s'agit du processus de l'intégration d'une expression du discours dont les éléments sont libres dans le système de la langue, en en faisant un syntagme dont les éléments sont indissociables. Ces éléments perdent leurs sens propres pour constituer une nouvelle unité lexicale, autonome et à sens complet, indépendamment de ces composants.[1]Ainsi notre thème correspond, aux expressions dites figées , locutionnelles ou idiomatiques qui se définissent par les contraintes limitant leur morphologie à cause du blocage de certaines propriétés syntaxiques, et par la non-compositionnalité de leurs composants sémantiques, ou ce que nous appellerons *l'opacité sémantique* .Ces expressions sont de fait, une donnée de base incontournable, dans la description des langues, Elles représentent un fait hautement économique pour le fonctionnement du système, mais problématique pour les descriptions disponibles et les méthodologies de traitement en vigueur, ce qui lui accorde une dimension heuristique certaine.

C'est en partant d'un bilan global des études portant sur ce phénomène que nous allons essayer d'étudier la fréquence et les occurrences des expressions figées dans des textes journalistiques en utilisant les outils du TAL (*Traitement Automatique de la langue)* ,notamment ceux développés par le LADL (*le Laboratoire d'Automatique Documentaire et Linguistique*) . Nous allons voir, donc à travers cette étude comment nous pouvons décrire l'emploi et les caractéristiques des expressions figées dans les textes journalistiques au moyen du traitement automatique, tout en avançant l'hypothèse que les textes

[1] Il s'agit de la synthèse de la définition de l'entrée *« figement »* du *dictionnaire d'analyse du discours* de CHARAUDEAU Patrick et MAINGUENEAU Dominique,. Le Seuil , Paris, 2002, P. 262 , celle du *dictionnaire de linguistique* , Larousse et celle du *Dictionnaire de Linguistique et des Sciences du langage ,1994* .

journalistiques se caractérisent par un emploi important et marquant de ces expressions par rapport à d'autres genres de textes.

Choix du corpus

Pour étudier les expressions figées, nous avons choisi comme corpus des textes journalistiques puisque nous estimons que ce phénomène est bien apparent dans ce genre de textes. Ce corpus est constitué de différents numéros du journal algérien El Watan que nous avons récupéré de l'archive électronique du journal. Nous avons opté pour ce type de textes en raison de leur variété discursive et leur richesse lexicale. En fait le texte journalistique jouit de plusieurs éléments qui favorisent sa validité comme un champ d'étude pour le phénomène du figement.

D'abord, la détection de ce phénomène exige un corpus copieux et évolutif, c'est-à-dire, des centaines, voire des milliers de pages d'un discours produit dans des périodes consécutives. Alors il n'y aurait pas mieux qu'un journal publié quotidiennement. Ensuite le journal se caractérise par la variété , ainsi sur le plan thématique qu' au niveau des registres et des spécialités linguistiques surtout pour les journaux généralistes où on alloue une rubrique à chaque domaine (politique ,économie , société , culture et littérature , science et technologie ,sport et loisirs etc.). Ce genre de texte jouit également d'une *hétérogénéité énonciative* pertinente et représente par excellence l'acte polyphonique, dans la mesure où la prise en charge de l'énonciation est d'une altérité explicite *(hétérogénéité montrée*) et implicite (*(hétérogénéité constitutive*)[1]. Cette propriété fait que le texte laisse mettre en œuvre les énoncés et les mots de l'autre, à force d'être mémorisés et réemployés par les locuteurs, comme étant une partie du système de la langue ; ce qui correspond à ces expressions que l'on considère actuellement comme figées ou en cours de lexicalisation. En fin la disponibilité de ce genre de texte notamment en version informatisée, surtout pour notre étude où nous allons procéder à un traitement automatique.

Choix de la méthode

Nous allons entamer cette recherche en procédant à une approche notionnelle qui tend à investir le champ conceptuel des notions de *figement* et des *expressions figées* en s'appuyant essentiellement sur les travaux de Gaston Gross, notamment son ouvrage *les expressions figées en français* (1996) , ainsi que les travaux de Ruth Amossy et de Anne

[1] D'après Dominique Maingueneau, l'hétérogénéité correspond à l'entremêlement de divers types de séquences textuelle , de registres de langues, de genre de discours et surtout à la présence implicite ou explicite de discours d'autres – c'est-à-dire attribuable à une autre source énonciative . Quant à la distinction montée/ constitutive, elle est faite par J. Authier-Revuz en 1982.

Herschberg Pierrot sur les stéréotypes et les clichés ,et *le dictionnaire des locutions idiomatiques françaises* de Bruno Lafleur. De ce fait notre étude sera, en premier lieu une approche linguistique car le figement implique certaines questions de la linguistique générale telles que l'arbitraire du signe et la référence. Cette tendance est affirmé par Salah Mejri en disant que :

> *« - Le figement a une valeur heuristique certaine puisque son étude permet de reprendre des questions fondamentales de la linguistique générale comme l'arbitraire du signe, sa linéarité, la conceptualisation, la référence, etc. ;*
> *- Le figement, de par son origine discursive, conduit à reposer en des termes nouveaux la dichotomie saussurienne langue / parole, et par conséquent, la vision générale qu'on a de la linguistique, de son objet et de sa méthodologie »*[1]

Nous allons, par la suite procéder à une analyse lexico- syntaxique qui nous permettra de décrire les propriétés des ces expressions telle que *la polylexicalité* et *l'opacité sémantique* et de définir les critères qui président au fonctionnement du phénomène du figement, tel que *le blocage des propriétés transformationnelles*. En effet Le figement est loin d'être un fait linguistique accidentel ; il est au contraire une caractéristique inhérente aux langues naturelles car il occupe une place privilégiée parmi les procédés et les processus à l'oeuvre dans le renouvellement du lexique en touchant à tout le spectre catégorial des parties du discours et il assure, de plus la formation de certains outils syntaxiques : (déterminants complexes, locutions prépositionnelles et conjonctives etc.) .

Pour approcher notre corpus, nous avons opté pour les méthodes du TAL comme étant l'outil le plus approprié pour mener une investigation dans ce domaine. En effet la détection du phénomène du figement exige un corpus volumineux dont le traitement manuel serait une tâche infaisable. Par ailleurs l'explosion du nombre des textes électroniques disponibles, surtout sur Internet a rendu, depuis quelques années le traitement automatique des langues naturelles et ses applications incontournables dans un grand nombre de recherches linguistiques[2].

[1] MEJRI, Salah, *Le figement, nouvelles tendances* , Université de Manouba, Tunisie , 2003, p.2.
2 CONSTANT, Matthieu, *Vers la construction d'une bibliothèque en-ligne de grammaires linguistiques*, Université de Marne-la-Vallée, http:// www.ladl.univ-mlv.fr , 2002,p.01, consulté le 16/06/208.

.Ces méthodes commencent avec les travaux du Laboratoire d'Automatique Documentaire et Linguistique (LADL), fondé en 1967 par Maurice Gross, qui se proposent de fournir, de manière systématique une description des expressions figées, aussi bien d'un point de vue syntaxique que sémantique. Cette approche était à vocation distributionnelle et transformationnelle en s'appuyant sur les travaux de Z. Harris qui pense que les phrases élémentaires ou (noyaux) constituent l'unité de base de la composition syntaxique et elles représentent les unités sémantiques de base et non pas les mots[1].

La méthode que nous avons choisi pour reconnaître et analyser l'emploi des expression figées dans notre corpus consiste à appliquer des ressources lexicales électroniques au moyen d'un logiciel de traitement automatique des textes .Le ressources lexicales en TAL sont les dictionnaire électroniques, les tables de lexique grammaire et le transducteurs ou les grammaires sous formes de graphes ; nous allons utiliser, particulièrement pour notre application les dictionnaires DELACF qui comporte les mots composés et la table C1d des expressions figées . Pour ce faire nous avons opté pour le logiciel *INTEX* créé au LADL en 1993 et développé en un environnement de développement linguistique par Max Silberztein en 2003 qui prend, actuellement le nom de *NOOJ*. Ensuite nous allons procéder au moyen de l'application *Nooj*, à une étude statistique et typologique des résultats fournis par le logiciel *Intex*. Cette étude nous permettrait de reconnaître et décrire la manifestation du phénomène du figement dans notre corpus de textes journalistiques.

Le plan

Nous étudierons dans le premier chapitre, la notion du figement dans la langue. Nous présentons, tout d'abord les termes et les concepts qui correspondent à ce phénomène, puis nous étudierons l'apport d'autres notions et disciplines dans le champ conceptuel de ce phénomène, tels que les clichés, les stéréotypes, les topoi et la rhétorique. Ensuite, nous analysons, plus en détaille cette notion en expliquant les définitions accordées aux termes *figement, expression figée et locution* .Enfin nous situons le figement dans le domaine de la composition, ce qui nous conduit à déterminer les deux contraintes principales qui président à la définition du figement, à savoir la restriction syntaxique et l'opacité sémantique.

[1] GROSS, Maurice, , « *Les limites de la phrase figée* », Langages, Larousse, Paris. n°90, 1988, p47.

Nous consacrons le deuxième chapitre aux études faites sur les critères du figement. Nous commençons par démontrer l'importance de la détermination de ces critères pour la définition et la délimitation de la notion et des catégories des expressions figées. Ensuite, nous exposons les critères proposés par Hudson et nous analysons plus en détaille ceux qui sont adoptés par Gaston Gross. Nous faisons, enfin la lumière sur les études de Maria Svensson, notamment son ouvrage intitulé « *critères du figement* » et nous analysons, plus particulièrement la notion de la mémorisation proposée par cet auteur.

Quant au troisième chapitre qui s'intitule *approches linguistiques et enjeux informatiques*, nous y menons une investigation sur les approches des expressions figées dans les différentes disciplines de la langue, notamment la linguistique traditionnelle, la sémantique et la lexicographie. Ensuite, nous étudions l'approche du traitement automatique de la langue (TAL) en insistant sur les travaux du LADL, surtout ceux menus par Maurice Gross qui , influencé par les linguistes américains Noam Chomsky et Z. Harris, a lancé un vaste programme de description systématique des propriétés syntaxiques de tous les éléments du lexique français. Nous étudions, également l'approche du lexique-grammaire ou lexique syntaxique adopté par Maurice Gross et son équipe de recherche qui vise à étudier systématiquement les propriétés syntaxiques de tous les mots du lexique commun en examinant toutes les constructions dans lesquelles entre chaque mot et les transformations qu'elles pourraient subir. Le LADL a consacré une partie importante de ses études aux constructions figées et aux contraintes qu'elles présentent en élaborant des tables décrivant les propriétés morphosyntaxiques de ces séquences qui sont dénommées *les tables de lexique-grammaire*. De ce fait, nous présentons, dans cette même partie quelques méthodes et outils pour la reconnaissance des expressions figées qui s'appuient sur l'analyse automatique des fréquences des suites de plus d'un mot dans un corpus informatisé . Nous décrivons, en l'occurrence les applications: Dic Assist, ACABIT et LEXTER , et nous signalons l'importance des méthodes qui se basent sur l'application des ressources lexicales, notamment les dictionnaires électroniques et les tables de lexique grammaire .

Dans le dernier chapitre, nous passons à l'analyse de notre corpus en appliquant les méthodes du TAL . Nous avons opté pour la méthode des ressources lexicales, notamment l'application des dictionnaires électroniques constitués de tables de lexique-grammaire décrivant des expressions figées et des mots composés. Pour ce faire, nous aurons recourt à deux applications informatiques. Nous présentons tout d'abord notre corpus constitué de différents numéros du journal El Watan. Nous expliquons, également les notions des

dictionnaires DELAF, tables de lexique-grammaire et transducteurs (grammaire sous forme de graphe). Ensuite, nous commençons par l'application Intex, qui est un logiciel conçu et développé au sein du LADL par Max Selberztein ; nous expliquons, en l'occurrence les fonctionnalités et le fonctionnement de cet outil de traitement automatique des textes notamment ceux qui concernent la reconnaissance et l'analyse des expressions figées au moyen des tables de lexique-grammaire et des transducteurs.

Dans la section suivante , nous passons à l'application de ce logiciel sur notre corpus, ce qui nous permettra, en premier lieu de relever des informations statistiques du corpus comme le nombre des phrases , de lexèmes, et des formes simples etc. . En suite, nous effectuons au moyen de cette même application, l'analyse lexicale du corpus en appliquant les ressources disponibles sur ce programme notamment la table de lexique grammaire C1d des locutions verbales et les dictionnaires des mots composés. Nous retenons, par la suite les résultats qui servent les objectifs de notre recherche, il s'agit de la liste (Frozen expressions) contenant les locutions verbales figées et la liste (compound words*)* des mots composés qui comportent les noms composés, les locutions adverbiales, prépositives, conjonctives et les pronoms composés. Enfin et afin d'étudier la typologie et la proportion de ces séquences dans le corpus nous analysons les résultats obtenus, à l'aide d'une autre application informatique appelée *Nooj*, laquelle se veut un environnement de recherche linguistique automatique développé à partir du logiciel *Intex*. Par ailleurs nous nous servirons de cet outil informatique pour démonter la particularité de la présence des expressions figées dans les textes journalistiques en comparant les occurrences de ces séquences dans ce genre de texte à leurs occurrences dans un texte romanesque .

Le figement dans la langue

1. Introduction

L'étude des expressions figées s'intéresse d'une propriété des langues naturelles dont l'importance a été méconnue pendant longtemps .il s'agit du phénomène du figement qui n'a pas été totalement ignoré mais son ampleur échappait à la plupart des auteur. Parmi les premiers qui ont évoqué ce phénomène on cite O. Jespersen qui suppose l'existence de deux principes opposés dans la langue : liberté combinatoire et le figement .Cette conception qui était une innovation dans le domaine des sciences du langage a été suivi par plusieurs études, cependant ces travaux ne sont pas entrées dans les écoles et n'ont pas franchi la barrière des programmes scolaires .Ainsi, il y avait une perception collective simpliste du mot composé comme cette forme qui a trait d'union . Le phénomène du figement a été occulté par l'absence de dénominations conventionnelles et de définitions rigoureuses de sorte qu'on est en présence de strates définitionnelles très souvent incompatibles.[1]

2. La terminologie du figement

Gaston GROSS nous explique que la confusion terminologique qui règne dans ce domaine due à deux raisons majeures :

2.1. Les différents aspects et dimensions perçus du figement

Pour illustrer cette question nous allons analyser les définitions d'un ouvrages qui fait autorité : Le dictionnaire de linguistique (Larousse) :

> « *Le figement est un processus linguistique qui, d'un syntagme dont les éléments sont libres fait un syntagmes dont les élément ne peuvent être dissociés. Ainsi, les mots composés (compte rendu, pomme de terre, etc.) sont des syntagmes figés. »*[2]

Nous remarquons ici que cette définition ne prend en compte que les syntagmes et leur passage de la liberté au figement et passe sous silence d'autres entités comme les déterminants, les adverbes, les prépositions et phrases et d'autres aspects comme la sémantique de ces suites.

[1] GROOS, Gaston., *Les Expressions figées en français*, Ophrys, France, 1996, 3.
[2] DUBOIS J. et al. *Dictionnaire de linguistique*, Larousse, Paris, 1973.

Cependant, ce même ouvrage définit le terme idiomatique en insistant sur l'aspect sémantique des expressions figées.

> « *On appelle expression idiomatique toute forme grammaticale dont le sens ne peut être déduit de sa structure en morphèmes et qui n'entre pas dans la constitution plus large : Comment va-tu ? How do you do ? sont des expressions idiomatiques.* » [1]

Alors le figement est pris, de ce terme et cette définition du point de vue sémantique en évoquant la non- compositionnalité du sens et les aspects syntaxiques sont à leur tour négligés ? Gaston Gross critique cette définition en disant : « *On ne perçoit pas clairement pourquoi on affirme que dans la phrase (comment va-tu) le sens n'est pas compositionnel* »

Passons maintenant à une autre définition qui relève du même domaine .le terme idiotisme est défini dans ce même dictionnaire comme suit :

> *« On appelle idiotisme toute construction qui apparaît en propre à une langue donnée et qui ne possède aucune correspondance syntaxique dans une autre langue Le présentatif c'est est idiotisme un gallicisme propre au français ; How do you do est un anglicisme »*[2]

On voit que l'aspect du figement a de nouveau changé. Cette fois on ne s'intéresse plus au figement sémantique ou syntaxique mais à l'impossibilité de traduction terme à terme d'une langue à une autre.

Par ailleurs, d'autre terme comme « mot composé » implique par sa définition un autre aspect du figement qui est l'aspect de la morphologie .En effet Ce terme est ainsi défini

> *« On appelle mot composé un mot contenant deux où plus de deux morphèmes lexicaux et correspondant à une unité significative : chou-fleur, malheureux, pomme de terre sont des mots composé. »*[3]

Cette notion tend à opposer la composition à la dérivation,ce qui n'est pas toujours pertinent .Le caractère flou et contradictoire de ces définitions apparaît clairement quand on s'aperçoit que le mot malheureux est classé, tantôt composé tantôt comme dérivé.

[1] DUBOIS J. et al. Op.cit.
[2] Ibid.
[3] Ibid.

Ces constations faites par Gaston GROSS l'ont amené à justifier cette confusion par l'absence de critères précis et en nombre déterminé, en vertu desquels on peut définir le figement.

2.2. Terminologie propre à certains linguistes

Un autre élément s'implique dans cette situation c'est la terminologie particulière propre à certains linguistes puisant de références théoriques différentes .E. Benveniste oppose un nouveau terme *synopsie* au mot composé et au mot dérivé et il attribue à ce terme la définition habituelle du mot composé : unité significative composée de plusieurs morphèmes lexicaux.

André Martinet, dans un article intitulé *Syntagme et Synthème,* paru en 1967, à introduit le terme *synthème* qui comprend selon lui : une séquence formé de plusieurs monèmes lexicaux fonctionnant comme une unité syntaxique minimale. Cette notion comprend aussi des mots dérivés comme *désirable, refaire* au même titre que les mots composés.

B .Poitier (1987) opte pour d'autres termes *Lexie composée* qu'il définit comme un ensemble comprenant plusieurs mot intégrés : *Brise-glace* et *lexie complexe* qui est pour lui une séquence figée comme *faire une niche, en avoir plein le dos*. Poitier considère la structure *avoir peur* comme séquence figée ce qui Gross réfute *« puisque peur est une substantif prédicatif que le verbe avoir est un verbe support qui peut être effacé après la formation de relative : Luc a peur, la peur que Luc a »*[1] .On note, aussi que la définition de la lexie ne montre pas si le figement est syntaxique seulement où aussi sémantique.

Dans ce même sens, Marie Véronique Le Roi, au cours de sa recherche sur les locutions verbales figées évoque cette idée sous l'intitulé de *profusion terminologique*. Dans cette section, elle affirme que, pour décrire le figement les auteurs ont proposés différents termes qui leur permettent d'exprimer des nuances dans leur théorie du figement.

Elle commence par certains auteurs classiques qui ont proposé de différentes dénominations illustrant, en fait les points de vue théoriques divergents et elle insiste sur l'idée que le figement est un phénomène irrégulier. Ainsi, Ferdinand de Saussure parle dans ses fameux *Cours de Linguistique Générale* (1916) d'*expression ou de*

[1] GROSS, Gaston, Op.cit, p. 5

locution toute faite. Cette qualification laisse transparaître le caractère immuable inhérent à ce type d'expressions:

> *« Le propre de la parole, c'est la liberté des combinaisons. Il faut donc se demander si tous les syntagmes sont également libres. On rencontre un grand nombre d'expressions qui appartiennent à la langue ; ce sont les locutions toutes faites, auxquelles l'usage interdit de rien changer […]*. »

Ensuite, son disciple Charles Bally consacre, dans son *Traité de stylistique* un chapitre aux locutions phraséologiques. Il distingue deux types de ces locutions :

a- Les *séries* phraséologiques qu'il a ainsi définies :

> *« Les éléments du groupe conservent leur autonomie, tout en laissant voir une affinité évidente qui les rapproche, de sorte que l'ensemble présente des contours arrêtés et donne l'impression du '' déjà vu''. »*[1]

Charles Bally parle ici d'une affinité et d'un rapprochement entre les éléments des groupes de mots qui assure une cohésion relative des termes et il compare ce phénomène à l'impression de *déjà vu*. Ce phénomène serait d'après Véronique Le Roi le figement dont parleraient d'autres linguistes ultérieurement.

b- Les *unités* phraséologiques sont les locutions où la cohésion des termes est absolue et qu'il définit comme suit :

> *« Une unité phraséologique représente un groupe de mots où «les mots qui composent le groupe perdent toute signification et l'ensemble seul en a un. […] Cette signification doit être nouvelle et non équivalente à la somme des significations des éléments. »*[2]

Charles Bally désigne par le terme unité phraséologique les expressions figées proprement dites. En effet, il parle d'une cohésion *absolue* des termes. Néanmoins, Véronique Le Roi est d'avis que Bally privilégie le critère intuitif pour la distinction des locutions phraséologiques.

Henri Frei évoque l'idée de la *brièveté sémantique* qu'i désigne sous le terme de *brachysémie* ou figement. Il le définit dans son ouvrage *Grammaire des* fautes de la manière suivante :

> *« Le mécanisme de la brachysémie ou brièveté sémantique est le figement d'un syntagme, c'est-à-dire d'un agencement de deux ou plusieurs*

[1] BAILLY, Charles, *Traité de stylistique française, Librairie Georg*, Paris, 1951.
[2] Ibid

signes, en un signe simple. La brachysémie, brièveté sémantique se distingue de la brachylogie, brièveté formelle. »[1]

Véronique Le Roi revient sur les travaux de Emile Benveniste avec plus d'illustration en distinguant trois types de formes complexes expliquées dans le tableau suivant :

Termes	Définitions	Exemples
1. Composés	*Unités à deux termes identifiables par le locuteur*	*Portefeuille*
2. conglomérats	*Unités nouvelles formées de syntagmes complexes comportant plus de deux éléments*	*Va-nu-pieds* *Meurt-de-faim*
3. Synopsie	*Groupe entier de lexèmes, reliés par divers procédés formant une désignation constante et spécifique*	*Fusil de chasse*

Tableau.I.1.Les trois différents types d'unités complexes selon E.Benveniste.[2]

Ce qui nous intéresse le plus dans ce tableau c'est le terme *synopsie* que E.Benveniste propose pour désigner des séquences de mots présentant un caractère figé et pour mettre en évidence le fait qu'il s'agit d'un modèle de construction différent de celui de la composition classique. En effet, Il distingue la synopsie des mots composés et des mots dérivés. Ainsi, *machine à coudre* est une synopsie, *timbre-poste* est un mot composé et *ferblanterie* est un mot dérivé.

Véronique Le Roi illustre, aussi la thèse de. Bernard Pottier dans son ouvrage *Linguistique Générale*, et *Introduction à l'étude des structures grammaticales fondamentales* (1962), qui distingue trois type de lexies :

Terme	Exemple
Lexie simple.	Cheval.
Lexie composée.	Cheval-vapeur.
Lexie complexe	Cheval marin.

Tableau.I.2.*Les trois différents types d'unités lexicales selon B. Pottier*[3]

[1] LE ROI, Véronique -Marie, *Traitement automatique et lexicographique des locutions verbales figées en françai*s , mémoire soutenu à l'université Paris III Sorbonne nouvelle ILPGA , p.13. www. cavi . univ-paris3.fr/Ilpga/ilpga/tal/sitespp/maitrise-2004/slMVLeroi-2004.pdf.
[2] Ibid, p.13.
[3] Ibid, p.14

Le figement concerne la lexie composée et la lexie complexe ; cependant la notion d'expression figée concerne beaucoup plous la lexie complexe qui désigne une séquence en voix de lexicalisation à des degrés divers. En effet, il s'agit d'une séquence qui peut être figée ou non ; selon le critère de séparabilité qui permet de les reconnaître et de les distinguer des syntagmes où les éléments du groupe sont séparables.

En 1985, Maurice Grosse emploie la qualification figée pour désigner certaines phrases, en fait il parle de *phrase figée* dans le cadre de ses travaux qui s'inscrivent dans la théorie de lexique-grammaire. Dans son approche, la phrase constitue l'unité sémantique de base ; les mots ou les morphèmes ne sont donc pas dans le cadre de ses travaux les unités minimales. Il n'est donc jamais question de *locution* pour désigner les séquences figées tout comme il n'est jamais question de *syntagme* pour se référer aux séquences libres.

3. Locution où expression ?

Revenons maintenant à la première partie de L'intitulé de notre recherche qui est *expression* et dont l'emploi est autant confondu que celui de *locution* surtout dans le domaine du figement. Bruno Lafleur parle de cette similarité entre ces deux terme et de leur définition générale comme un groupe de mots qui exprime une chose, une action où une idée mais il affirme que *« la nuance est bien mince entre locution et expression*[1] » .Cependant on pourrait distinguer le terme locution par son emploi pour des qualifications correspondant aux différentes classe grammaticales : *rapidement* est verbe , *à la hâte* est une locution adverbiale ; *se chicaner* est verbe , *avoir maille à partir* est locution verbale ; *avar* est un adjectif tandis que *dur à la détente* est locution adjectivale .

Nous remarquons que la confusion ne se fait pas entre locution et expression , mais entre locution et expression figée . ça se manifeste dans l'intitulé de l'ouvrage de Gaston Grosse *Les expression figées en français* qui est sous titré comme suit : *Noms composé et d'autres locutions* cette mention nous fait déduire que les locutions représentent une sous-catégorie des expressions figées donc, d'après Gross les locutions sont des expressions figées et les définit comme « *tout groupe dont les éléments ne son pas actualisés individuellement* »[2] .

[1] LAFLEUR, Bruno, .*Dictionnaire des locutions idiomatiques françaises*. Duculot, Ottawa, 1979, p.v.
[2] GROSS, Gaston. Op.cit, P. 14

L'expression *un cordon bleu* est considéré comme une locution où expression figée parce que son sens n'est pas constitué de sens des éléments qui la composent tels qu'ils fonctionnent en de hors de cette séquence ,qui a un sens tout a fait indépendant désignant (quelqu'un qui fait bien la cuisine) .En plus de leur figement sémantique les locutions présentent un figement syntaxique : on ne peut pas insérer d'adverbe devant l'adjectif *bleu un cordon assez bleu** et on ne peut pas lui substituer un synonyme .En revanche cette expression fonctionne comme unité significative autonome *un excellent cordon bleu*

4. Le figement et les clichés

Certains auteurs appellent ces locutions des clichés et ils s'accordent à dire qu'ils s'agit d'expressions toutes faites, de banalité, de lieux communs mais Bruno Lafleur ne voit pas que des expression : cheveux d'or, lèvres vermeille, teint de rose, l'aurore au doigts de rose, le blanc manteau de l'hiver, le feu brûlant de la passion sont des locutions idiomatiques M .Marouzeau définit le cliché ainsi.[1] :

> « *expression suffisamment typique pour être reconnue de prime abord , à laquelle recourt le sujet parlant et surtout l'écrivain soucieux d'imiter ce qu'il estime être une élégance , et qui souvent, à force d'être usée, donne l'impression de la pire banalité : jeter son dévolu, sombrer dans le marasme »*[2]

Par ailleurs, Lafleur opte pour la définition du maître Charles Bally comme il le qualifie qui dit : « *Les clichés sont des locutions toutes faites, transmises par la langue littéraire à la langue commune »*[3] .Donc le figement ici résulte de l'imitation des grands auteurs par les élèves, et les débutants et les mauvais auteurs comme dit Charles Bally : *le printemps de al vie, l'hiver des ans, l'astre du jour, la reine des nuits*. Cependant on peut généraliser parce que certaines locutions peuvent venir de la langue littéraire sans pour autant qu'elle soient considérées comme des clichés comme le cas des exemples suivants : *attacher le grelot, montrer patte blanche, appeler un chat un chat*, qui sont les deux premiers de la Fontaine et l'autre de boileau.

En résumé, nous pouvons dire que clichés et locution sont en intersection. C'est-à-dire toute les locutions ne sont pas des clichés (les locutions conjonctives où

[1] BRONO, Lafleur. Op.cit, p.5
[2] MAROUZEAU. Jean, *Lexique de la terminologie linguistique*, Librairie orientaliste Paul Geuthner, 1962, Paris
[3] BALLY, Charles, Op.cit.

prépositives et beaucoup de mots composés) mais certaines d'entres elles le sont : *une faim de loup* .Pour Ruth Amossy, les clichés correspondent en particulier à des expressions marquant l'intensité, fondées sur des comparaison *beau comme un dieu, une fièvre de cheval, un patience d'ange* ou des métaphores figées rouler à tombeau ouvert. d'autre part, tous les clichés ne sont pas des locutions car le clichés peut être des séquences qui ne présentent pas un figement mais il s'agit d'une simple fréquence dans un genre de discours comme la séquence *éminent linguiste* qui est considéré comme association clichée mais leurs éléments sont pourvus d'une certaine autonomie syntaxique .Nous notons que l'étude des clichés s'inscrit généralement dans une approche stylistique qui étudie leurs effets dans un contexte discursif , leur rôle dans la production du texte et les différentes lectures auxquelles ces associations peuvent donner lieu .

5. Le figement et les stéréotypes

Le stéréotype pourrait, également être considéré comme une forme du figement qui correspond beaucoup plus à la strate sémantique. En effet il a été abordé par le philosophe américain Hilary Putnam comme l'objet d'une théorie sémantique dans son article *Is semantics possible ?*en 1970 qui porte sur la signification des noms d'espèces naturelles. Selon Putnam, le stéréotype est une idée conventionnelle, associée à un mot dans une culture ou communauté donnée. Ainsi pour le tigre on associe les rayures, pour le citron, l'acidité et un type de peau épaisse, pour l'eau sans couleur, transparence, sans goût qui étanche la soif, etc. ». De ce fait le stéréotype est une expression employée pour associer à un mot un sens général lié à une culture ou une communauté données.

> *C'est une représentation simplifiée associé à un mot obligatoire pour assurer un bon usage de la communication (…). Le stéréotype assure une description du sens en usage, fondée sur la reconnaissance de la norme sociale et culturelle.*[1]

Cette idée s'oppose à la théorie des conditions nécessaire et suffisantes « le sens d'un mot, étant compris comme ce qui détermine sa référence, est constitué des

[1] AMOSSY, Ruth, HERSCHBERG PIERROT, Anne, *Stéréotype et clichés*, Arman colin, France, 2005, p .89.

conditions que doit remplir un référent pour être adéquatement dénommé par ce mot ».[1] En effet, des citron dont la peau n'est jaune sont toujours des citrons (la couleur jaune est un trait distinctif nécessaire mais non suffisant de la définition des citrons).

Le rapport entre le stéréotype et les locutions ou les l'expressions figées c'est que le stéréotype, comme étant un figement au niveau des idées attachées à des unités lexicales se manifeste dans le langage sous forme des expressions figées ou en cours de figement.

> « *Le stéréotype n'est pas, d'ailleurs contenu dans les définitions. Dans une étude comparé entre de l'espagnol et du français Ariane Desportes et Françoise Martin-Berthet (1995) souligne la nécessité de prendre en compte les unités phraséologiques de chaque langue pour décrire les stéréotypes »*[2]

Nous concluons que locution, cliché et stéréotype font partie d'un continuum des expressions figées, avec les adverbes *aide –toi, le ciel t'aidera* et les slogans *le poids des mots , le choc des photos.*

6. Le figement et les topoi dans la pragmatique intégrée

Ce terme d'origine aristotélicienne est redéfini dans le cadre de la théorie dite *l'argumentation dans la langue* édifié par Jean-Claude Anscombre et Oswald Ducrot .Dans cette théorie on pense que la force ou les valeurs argumentatives sont indissolublement liées à la signification du mot, de l'expression ou de l'énoncé .D'où vient l'idée de *la pragmatique intégrée* où le sens profond d'un énoncé ne peut pas être séparé de son utilisation en contexte et de sa valeur argumentative . De ce fait les topoi représentent, eux aussi une forme de figement sur le plan sémantique ce qui apparaissent clairement dans leur définition :

> *« Principe généreux qui servent d'appui aux raisonnement (…) ils sont presque toujours présentés comme faisant l'objet d'un consensus au sein d'une communauté plus où moins vaste »* [3]

Anscombre parle de consensus ce qui signifie des idées généraux communs « Présentés comme acceptés par la collectivité » (Ducrot 1988 : 103) pour assurer des liens conclusifs entre les énoncés appelés les topos. Ces principes et idées communs

[1] MARTIN-BERTHET Françoise, Définitions d'enfant : étude de cas, Repère, n°8, 1993, p.117.
[2] AMOSS, Ruth, HERSCHBERG PIERROT, Anne , Op. cit, p. 92.
[3] ANSCOMBRE, Jean-Claude, Théorie des topoi, Kimé, Paris, 1995, p. 36.

vont se réaliser sous forme des pratiques discursives communes — en l'occurrence des formes d'expressions figées.

7. le figement et la rhétorique

La plupart des locutions idiomatiques sont d'après Lafleur métaphoriques et c'est ce qui fait leur charme .Toute la gamme des figures de style et de rhétorique y passe, depuis la catachrèse, l'euphémisme, l'antiphrase, en passant par la comparaison, la litote, à quoi il faut ajouter les tournures intensives, affectives, l'onomatopées etc. Lafleur pense que ce n'est pas ces appellations et cette typologie qui permettent au lecteur et l'auditeur de comprendre et sentir l'effet des ces figures mais il est pour l'avis de M.Henri Morier qui réclame une analyse psychologique :

> *« La rhétorique moderne ne devrait pas se borner aux indications qui permettent de fabriquer une figure ; elle devrait surtout l'étudier du point de vue psychologique, voir ce que se passe dans l'âme du lecteur au moment ou la figure y pénètre, comment est elle s'y décompose en développant une énergie qui émeut la sensibilité, en un mot comment elle agit pour causer en nous cet émerveillement qui est un effet de l'art "*[1]

En effet les locutions figées comportent, généralement des figures de style qui demeurent, malgré leur vieillesse encore plus évocatrice d'émotion et les meilleurs écrivains ne craignent pas de les employer surtout lorsqu'elles se présentent naturellement sous leurs plumes , ce n'est pas suite d'une pauvreté de vocabulaire ou par manque d'imagination mais c'est une affaire de style encore plus que de langue .Donc le figement ne porte pas seulement sur la langue mais il correspond aussi au style et à l'image . [2]

8. Le figement et les registres de langue

Certains dictionnaires qualifient ces locutions de familiers ou populaire en obéissant à de vieux perception .Charles Bally a écrit au début du siècle que " porter quelqu'un aux nus "est extrêmement familier Bruno Lafleur se demande ici si Charles Bally tiendrait encore ,aujourd'hui à cette qualification. En effet la perception des expressions figées passe, au fil de temps d'un registre à un autre.

[1] MORIE, Henri, *Dictionnaire de poétique et de rhétorique*, p.VII
[2] LAFLEUR, Bruno, Op.cit, p. VIII.

Quant à la langue écrite, on peut la qualifier de didactique, techniques, administratives ou académique et tout le reste, selon Lafleur est littérature. Pour chacun de ces registres on peut associer un certain nombre d'expressions figées qui marquent le discours dans ce domaine .De ce fait nous avons poser l'hypothèse que le discours journalistique serait marqué à leur tour par un certain nombre ou types d'expressions figées .

9. La notion du figement

Depuis longtemps, on considère le figement comme un phénomène linguistique complexe et irrégulier et qui a donc un caractère marginal dans la langue. Maurice Gross a pourtant expliqué dans ses travaux que d'un point de vue statistique ces sentiments d'irrégularité et d'exception n'avaient pas lieu d'être. En effet, il existerait près de 1800
constructions verbales qui ne mettent pas en jeu un emploi spécifique du verbe. C'est le cas
d'une phrase telle que :
- *Luc lèche le plat.*

Cependant, on trouve que 8000 constructions verbales seraient figées. L'exemple suivant montre bien que le verbe *lécher* n'est pas utilisé avec le même emploi que précédemment :
- *Luc lèche les bottes de Max.*

M. Gross estime donc qu' « *ignorer ces constructions revient à ignorer une bonne partie du* langage ».[1]

Le phénomène de figement a été abordé par ailleurs par Otto Jespersen qui était un précurseur dans ce domaine. Ce linguiste a distingué, dans son ouvrage *Philosophy of Grammar* en 1924, deux principes dans les langues : la liberté combinatoire et le figement. Cette manière d'aborder les langues accorde un caractère essentiel au processus de figement.

Weinrich accordait aussi une grande importance aux expressions figées. Il disait à propos du figement : « *Ce qui avait longtemps été considéré comme un phénomène marginal, comme une série d'exceptions, se révèle être en fait caractéristique des langues humaines naturelles* ». Gaston Gross surenchérit en accordant la même

[1] GROSS, Maurice, *Les limites de la phrase figée*, Langages, n° 90, Larousse, Paris, 1988.

importance au phénomène des expressions figées qu'à la double articulation d'André Martinet.

En outre, certains auteurs ont tendance à accorder une trop grande importance au phénomène en disant que tout est phraséologique. Alors, Les nombreuses et diverses définitions et dénominations qui ont été introduites par les différents auteurs et leurs ouvrages pour décrire ce même phénomène ont contribué à apporter au figement un caractère marginal et irrégulier.

9.1 Définitions

De différentes définitions sont proposées pour le nom *figement* ou l'adjectif *figé*. Nous présentons ici certaines de ces définitions extraites de divers dictionnaires et ouvrages et citées par Marie Véronique le Roi dans son mémoire de recherche intitulé *Le traitement automatique et lexicographique des locutions verbales figées en français*.

On commence par le dictionnaire *Lexis* qui définit l'adjectif *figé* comme suit: « *se dit d'un mot, d'une construction qui cessent de subir dans la langue une évolution.* »

Le Petit Robert définit la qualification *expression ou locution figée* ainsi*: « dont on ne peut changer les termes et qu'on analyse généralement mal* ».

Nous remarquons que ces définitions sont pour le moins laconiques et se contentent de souligner l'existence du phénomène tout en supposant que celui-ci est irrégulier.

Alors que Alain Rey et Sophie Chantreau fournit davantage de précision, dans leur *Dictionnaire d'expressions et locutions* en 1997:

> « *Un lexique ne se définit pas seulement par des mots simples et complexes, mais aussi par des suites de mots convenues, fixées, dont le sens n'est guère prévisible […]. Ces séquences, on les appelle en général des locutions ou des expressions.* »[1]

D'autres précisions sont données dans les dictionnaires de linguistique.

Le Dictionnaire de linguistique Larousse en s'appuyant sur des exemples donne une définition[2] que Marie Véronique Le Roi qualifie comme un peu moins vague et qu'elle commente comme suit : « *Nous verrons plus avant que la composition et le*

[1] REY, Alain, CHANTREAU, Sophie, *Dictionnaire d'Expressions et Locutions*, Dictionnaires Le Robert, Paris 1989.
[2] Voir supra, I, 2.1, p1.

figement sont des phénomènes distincts et que tous les mots composés ne sont pas nécessairement figés. »[1]

Quant au dictionnaire de Linguistique et des Sciences du Langage (1994), il définit le figement ainsi:

> *« Le figement est le processus par lequel un groupe de mots dont les éléments sont libres devient une expression dont les éléments sont indissociables. Le figement se caractérise par la perte du sens propre des éléments constituant le groupe de mots, qui apparaît alors comme une nouvelle unité lexicale, autonome et à sens complet, indépendamment de ses composants. »*[2]

Pour J.C. Anscombre le figement est un processus au terme duquel le locuteur n'est plus capable de déterminer le sens d'une séquence à partir de celui de ses constituants. Et Georges Misri (1987), à son tour, désigne sous le terme expression figée :

> *« Tout groupe de monèmes qui présente un blocage total ou quasi total des axes paradigmatiques et syntagmatiques, c'est-à-dire une impossibilité ou une réduction importante des possibilités de commutation et / ou d'expansion partielle.»*[3]

Nous signalons ici que ces différentes définitions tendent à montrer que le figement est un phénomène hors norme et irrégulier.

9.2. Le figement et la composition

Le phénomène du figement était abordé par les manuels classiques de lexicologie dans la partie traitant de la composition. De ce fait nous allons essayé, dans un premier temps de expliciter la corrélation entre le figement et la composition affin d'en déduire les caractéristiques générales des expressions figées.

Dans sa description du figement, Gaston Gross utilise un certain nombre de termes et de définitions spécifiques qu'il est possible de retrouver chez d'autres auteurs:

- Un groupe ou un syntagme est dit *libre* s'il correspond à une séquence générée par les règles combinatoires mettant en jeu à la fois des propriétés syntaxiques et sémantiques. L'adjectif *libre* s'oppose donc à l'adjectif *figé*

[1] LE ROI, Véronique-Marie, op.cit, p.11.
[2] Ibid, p.12.
[3] Ibid, même page.

-Un *idiotisme* (gallicisme, anglicisme ou germanismes) est une séquence que l'on ne peut pas traduire terme à terme dans une autre langue.

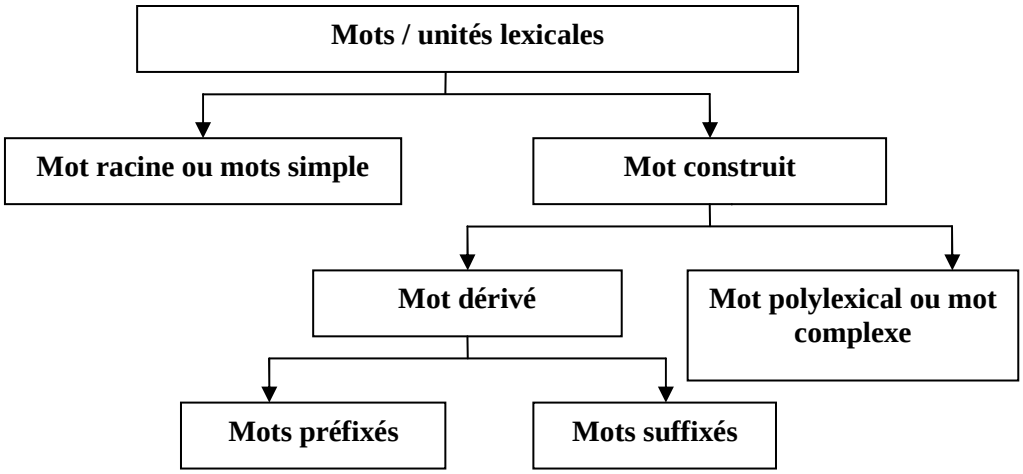
-Un *mot racine* ou un mot *simple* désigne toute unité qui n'est susceptible d'aucune décomposition.

- Un mot qui n'est pas un mot simple est alors dit *construit*. Les mots construits sont donc des mots composés de différents morphèmes autonomes. G. Gross distingue deux types de mots construits :

a. Les mots *dérivés* que l'on obtient par l'affixation d'un préfixe ou d'un suffixe à une base donnée ;

b. Les mots *polylexicaux* (ou mots complexes) qui désignent toute unité composée de deux ou plusieurs mots simples ou dérivés préexistants. Ces mots peuvent être soudés et donc ne pas comporter de séparateurs.[1]

Marie Véronique Le Roi illustre cette typologie de Gross à travers le schéma suivant :



Schema.I.1. Les différents types d'unités lexicales selon Gaston Gross[2]

La composition et la dérivation qui s'opposent dans ce tableau sont, donc les deux parmi les moyens de formation de nouvelles unités lexicales disponibles en français. La dérivation est un procédé récursif : la base à laquelle est affixée un préfixe ou un suffixe peut, elle-même être un mot dérivé. La composition est donc moins productive que la dérivation quant à la formation des nouveaux mots en français.

Figement et composition sont souvent confondus et considérés comme des synonymes. Mais cela n'est pas avéré ; en effet, une suite composée n'est pas

[1] GROSS, Gaston, Op. cit. , Chapitre II .
[2] LE ROI, Véronique-Marie, Op.cit, p. 16

nécessairement figée. Les suites composées peuvent être sémantiquement transparentes et à cet égard ne seront donc pas considérées comme figées.

Dans son étude sur le figement et notamment dans son ouvrage *les expressions figées en français* (1996), Gaston Gross met en évidence deux contraintes majeurs qui interviennent la description du figement :

> *« -Une contrainte d'ordre sémantique : l'opacité sémantique ; cette suite est-elle sémantiquement transparente ou opaque ?*
>
> *-Une contrainte d'ordre syntaxique : une suite donnée est-elle syntaxiquement libre ? »*[1]

En effet, le figement implique une opacité sémantique d'une part des contraintes, voir des blocages au niveau de certaines propriétés syntaxiques d'autre part. Par conséquent

une suite peut être considérée comme étant figée quand celle-ci est sémantiquement opaque et n'est pas libre syntaxiquement .

De fait, la notion du figement est étroitement liée à la notion de la composition dans la mesure où une suite est considérée figée quant celle-ci est lexicalement composée (*poly lexicale*[2]*)* et sémantiquement non-composionnelle.(*opaque*)[3]

[1] Ibid, p.16.
[2] Marie-Véronique définit le terme « *polylexical* » comme suit : une suite est dite« polylexicale » quand elle est composée de plusieurs éléments lexicaux qui ne jouent pas de rôle extérieur à la séquence. Autrement dit les éléments lexicaux contribuent uniquement à la constitution de la suite
[3] Voir infra II, 4.2, p.

II. Les critères du figement

1. Introduction

La définition des expressions figées nécessite la détermination de certains critères qui nous permettrons de définir l'expression figée comme catégorie. En effet, la notion d'expression figée comprend souvent différentes catégories (idiome, locution, proverbe, etc.).

Maria Svensson présente, dans son ouvrage *les critères de figement*, deux points de vue différents : le premier c'est celui de Hudson (1998 : 33) qui signale, dans le cadre des études sur les expressions figées anglaises, que « *les critères typologiques donnent souvent comme résultats des taxinomies incluant des catégories non discrètes et incomplètes.* »[1] (D'après la traduction de Svensson) ; le second c'est l'avis de Gaston Grosse qui insiste sur la description du phénomène du figement en se limitant à des catégories précises :

> *« Nous ne voulons pas réduire pour autant le figement aux considérations générales que nous proposons ici. C'est dans le cadre des différentes catégories que peut se faire l'analyse avec la précision voulue. »*[2]

Cependant, pour que l'on puisse délimiter des catégories des expressions figées, on doit, à priori déterminer les propriétés communes qui caractérisent l'expression figée, autrement dite les conditions nécessaires, pour attribuer une séquence à la catégorie des expressions figées. *Propriété* c'est le terme évoqué par Gaston Gross (1996) : « *Nous examinons les propriétés communes qui caractérisent ce phénomène que l'on doit considérer comme un des plus importants dans les langues.* ».[3] En revanche, d'autres auteurs comme S. Schabira (1999 :9) évoque le terme de critère que nous adoptons dans le cadre de cette recherche ; il affirme qu'il y a des critères distinctifs qui se vérifient pour toutes les expressions figées. Par ailleurs, il ajoute qu'il en existe d'autres critères moins rigoureux, qui s'appliquent ou non selon le degrés du figement plus ou moins élevé de la séquence et il explique, par la suite, que le figement est un phénomène scalaire, ses produits pouvant présenter un registre plus ou moins intense.

Par conséquent, *catégorie* et *critère* pourraient être considérés comme deux outils primordiaux pour l'analyse et la description du phénomène du figement. De ce fait,

[1]HUDSON, Jean, *Perspectives on fixedness: applied and theoretical*, Lund Studies in English, *94*, Lund University Press, Lund, 1998, p.33.
[2] GROOS, Gaston., Op.cit , p.9.
[3] Ibid.

Maria Svensson a consacré ses études, sur le figement, à ces deux notions, notamment les critères utilisés pour décrire les expressions figées, mais elle affirme que le terme n'est pas univoque ; en effet, il reçoit différentes acceptions :

> « *Tantôt il semble décrire ce qui est typique pour certains type d'expressions, tantôt il est utilisé pour, vraiment tranché entre les syntagmes figés et les syntagmes non figés* »[1]

2. Les aspects et les types des critères décrivant les expressions figées

2.1. 1es aspects couverts par les critères

> Pour bien expliquer la notion de critère Maria Svensson a essayé de cerner les aspects que couvrent ces critères en les résumant dans les points suivants : »

- Ce qui est typique pour toutes les expressions figées.
- Ce qui est typique pour certaines.
- Des traits qui séparent les expressions figées de la syntaxe libre.
- Des traits qui sont pertinents pour les expressions figées mais ils sont aussi valables pour la syntaxe libre.

La phraséologie connaît une abondance de terme employé dans la description des expressions figées hors que le terme critère. Ainsi les phrasiologues parlent de *propriété*, de *trait distinctif*, de *paramètres*, de *causes* voir de *symptômes*.

2.2. Les types des critères

Ces termes, d'après Maria Svensson, n'atteignent pas toujours le statut de critères distinctifs même, certains chercheurs les emploient les uns à l'exclusion des autres. Svensson parlent de deux types de critères :

a- critères appliqués dans les recherches phraséologiques tels que la non-composionnalité, la conventionalité, la métaphoricité, la mémorisation, l'inflexibilité syntaxique et syntaxe marquée. Svensson signale que ces critères ne sont pas toujours exprimés par ces termes mais ils couvrent à peu près les m^mes phénomènes entièrement ou partiellement.

[1]SEVNSSON, Maria Helena, *Le critères de figement .L' indentifications des expressions figées en français contemporain*, Print &Media, Umeá, 2004, p.27. www.duo.uio.no/roman/Art/Rf-16-02-2/fra/Svensson.pdf, consulté le 23/09/ 2007 .

b- D'autres critères qui sont les résidus de la langue ancienne et impliques des notions syntaxiques et lexicales marquée. Le caractère non officiel, la valeur intentionnelle ou la non –actualisation d'un référent, les restrictions sélectionnelles la non possibilité de traduire dans une autre langue etc.

3. les critères de Hudson

Par ailleurs, la notion du critère a été l'objet d'étude de plusieurs auteurs travaillant sur le phénomène du figement. Hudson (1998), dans sa thèse *Perspective on fixedness,* cite quatre critères récurrents dans les ouvrages traitant des expressions figées en anglais

1- Contraintes syntaxiques inattendues *« unexpected syntaxic contraints on the constuent parts »*
2- Restriction collocationnelle inattendues *« unexpected colloctioal restriction within thes expression »*
3- Syntaxe anomale ou usage anomal « *anamoloous syntax or usage »*
4- Sens fuguratif « figurative meanig »

Hudson insiste sur les premiers critères qu'elle appelle des critères *variationnels* « *Variabilité critiria* » et elle en fait la base pour sa définition du figement « *fixedness* » Cependant elle pense que ces critères ne sont que des syntagmes de sens figuratif et elle introduit un aspect que d'autres chercheurs oublient ou laisse de coté et qu'elle appelle les forces figéantes « fixing forces ».

4. Les études de Gaston Gross

Une autre étude qui se veut plus exhaustive. C'est celle de Gaston Gross (1996) dans son chapitre intitulé la notion du figement où il parle des propriété communes qui caractérisent le phénomène du figement .En effet Gaston Grosse entend étudier la notion générale du figement dans une perspective descriptive en vue de rendre compte de tous les aspects de ce phénomène. Pour ce faire Gross a énuméré onze propriétés à savoir : [1]

1. La polylexicalité.
2. L'opacité sémantique

[1] GROOS, Gaston., *Les Expressions figées en français*, Ophrys, France, 1996, 9,23.

3. Le blocage des propriétés transformationnelles.
4. La non- actualisation des éléments.
5. Portée du figement.
6. Degré de figement.
7. Blocages des paradigmes synonymes.
8. La non-insertion.
9. Le défigiment.
10. Etymologie.
11. Les locutions sont-elles réductibles à des catégories ?

Les éléments présentés par Gaston Gross ne pourraient pas être considérés tous comme des critères distinctifs mais il s'agit plutôt de notions plus au moins pertinents et utiles pour décrire le figement. Nous allons, par la suite rendre compte des quatre éléments que nous pourrions considérer comme critères distinctifs des expressions figées

4.1. La polylexicalité

La présence d'une séquence de plusieurs mots est, selon G. Gross une « condition nécessaire pour avoir affaire d'un figement et il exclut les suites qui résultent de la dérivation ; il exclut, aussi certaines restrictions morphologiques : « *Sur la base de l'adjectif gentil, on peut former un substantif dérivé à l'aide du suffixe d'état –ess mais non avec le suffixe –té, qui traduit, lui aussi, un état*.»[1] .Et il va de même pour d'autres restrictions syntaxiques quand chaque prédicat a un domaine d'argument qui lui est propre.

Il impose, ensuit une autre contrainte aux mots qui composent ces séquences. C'est que ces mots doivent avoir une existence autonome. En fin il évoque le cas des suites composées d'un élément latin(post-, anté-, extra-, sem-) ou grec (auto- ,hémi-, péri- etc.)et le problème des séparateurs entre les élément lexicaux (le trait d'union, l'apostrophe et le blanc) que l'on doit réduire à un simple problème de graphie. ET il se demande s'il faut accepter des séquences soudées comme vin et aigre et considérer le mot vinaigre comme une suite figée.

[1] GASTON, Gross, Op.cit, p. 9.

4.2. L'opacité sémantique :

Gaston Gross explique qu' une suite donnée peut avoir deux lecture possibles, l'une *transparente* et l'autre *opaque* De ce fait le concept de la compositionnalité de la grammaire traditionnelle n'est pas toujours valide « le sens d'une séquence est le produit de celui des éléments composant » Ainsi pour, la phrase « la moutarde lui monte au nez » le sens des mots qui la composent ne permet pas de conclure qu'il s'agit d'une personne qu' se fâche .Il va de même pour la phrase « *les carottes sont cuites* » qui, en plus de son sens non-compositionnel ou opaque qui signifie que « *la situation est désespérée* » elle peut avoir une interprétation compositionnelle pour dire que « *les légumes sont prêts* ».

Cette notion s'applique aussi aux unités de niveau inférieur. Ainsi le groupe nominal *clé anglaise* qui ne signifie pas une clé que l'on fabrique en Angleterre mais il s'agit d'un type particulier de clé. Donc le sens de cette séquence n'est pas déduit de celui ses éléments ; il n'est pas transparent par conséquent nous sommes en présence d'une opacité sémantique qui est d'après Gaston Gross une phénomène scalaire : « Elle peut être totale (La clé des champs), Partielle (clé anglaise), ou inexistante (clé neuve). » [1]

Maria Svensson souligne que Gaston Gross ne dit pas explicitement que l'opacité sémantique est une condition nécessaire pour le figement mais il affirme que l'opacité sémantique va de paire avec le figement syntaxique : « (…) *le figement sémantique et le figement syntaxique sont deux aspects d'un même phénomène qu'il convient de ne pas séparer de façon artificielle »*[2]. Cette idée s'oppose à l'opinion de Hudson qui fait une distinction claire entre ces deux aspects de la langue quand il explique son quatrième critère à savoir le sens figuratif.[3]

4.3. Le blocage des propriétés transformationnelles

Les séquences figées présentent une absence totale de possibilité de transformation par passivation, pronominalisation, détachement, extraction ou relativation. Toutes ces modifications et changements et de structures appelés « *transformation* » sont applicables pour une phrase comme *L'enfant a lu ce livre* suivant le tableau ci-dessus :

[1] Ibid, p.11
[2] Ibid, p. 8.
[3] Svensson, Maria, Op.cit, p. 29.

Passivation	*Ce livre a été lu par l'enfant.*
Pronominalisation	*L'enfant l'a lu.*
Détachement	*Ce livre, L'enfant l'a lu.*
Extraction	*C'est ce live que l'enfant a lu.*
Relativation	*Le livre que l'enfant a lu.*

Tableau.II.1.Les transformations possibles pour une phrase à construction libre[1]

Cependant, dans la phrase « *Luc a pris la tangente* » on ne trouve rien dans le verbe *prendre* ni dans le substantif *tangente* qui permet de prédire le sens opaque de cette séquence laquelle veut dire (se tirer d'affaire habilement, esquivé une difficulté). Cette opacité sémantique est, en fait corrélée à une absence de ces propriétés transformationnelle. Ainsi, les séquences du tableau suivant sont inacceptables.

Passivation	**La tangente a été prise par Luc.*
Pronominalisation	**Luc l'a prise.*
Détachement	**Cette tangente, Luc l'a prise.*
Extraction	**C'est cette tangente que Luc a prise.*
Relativation	**La tangente que Luc a prise.*

Tableau.II.2.Le blocage des propriétés transformationnelles des séquences transparentes (opaques)

Par ailleurs, ce blocage s'applique aussi au substantif. Ainsi, de diverses transformations, comme (la nominalisation, l'adjonctions d'adverbes intensifs, la prédicativité) sont interdites avec un groupe nominal composé d'un nom et d'un adjectif comme le cas des suites telles que *cordon (-) bleu* (bonne cuisinière), dont l'opacité sémantique est évidente. C'est ce que nous illustrons dans le tableau qui suit :

La nominalisation	*Le bleu de ce cordon.*
L'adjonction d'adverbes intensifs	*- Un cordon très bleu.* *- Un cordon particulièrement bleu.*
La prédicativité	*Ce cordon est bleu*

Tableau.II.3. blocage des propriétés transformationnelles des substantifs composés

[1] GASTON, Gross, Op.cit,, p. 12.

Gaston Gross insiste sur l'idée que ces restrictions syntaxiques sont bien liées à l'opacité sémantique et il conclut que :

> « *Le figement est un phénomène qui transcende les différents niveaux de l'analyse linguistique et qu'une description qui ne serait que syntaxique ou sémantique ne retiendrait q'une partie des faits.* »[1]

Dans le même sens, Maria Helena Svensson parle du critère de l'inflexibilité qui, en plus des blocages des propriétés transformationnelles évoqués par Gross correspond à des contraintes au niveau du changement de genre, de nombre ou de temps.

> « *Ainsi nous avons affaire à un figement si on ne peut pas changer genre nombre ou temps, comme dans rose trémière*[2]*, faire la loi - *faire les lois, les carottes sont cuites - *la carotte est cuite et qui vivra verra -*qui vit voit. Dans le cadre de l'inflexibilité, la possibilité d'effectuer des transformations, telle que pronominalisation, passivation et relativation, est aussi pertinente. »* [3]

4.4. La non-actualisation des éléments

Gaston Gross présente un troisième critère qui prolonge l'opacité sémantique et le blocage des propriétés transformationnelles. Il s'agit qu'aucun des éléments lexicaux constitutifs de la séquence ne peut être actualisé. L'actualisation permet d'inscrire un prédicat dans son contexte. Ainsi, un verbe est actualisé quand il est conjugué. Gaston Gross considère que la non-actualisation est critère principal et condition nécessaire pour définir une locution. Ainsi, dans la phrase *Paul nous a dit cela avec le désir de nous convaincre*, la locution prépositive *avec le désir de* est un prédicat qui a perdu son actualisation. En effet, cette locution se veut une instance intermédiaire entre une catégorie simple « *pour* » et le prédicat conjugué « *il avait le désir, il désirait ou il était désiré de* ». Dès lors, Gaston Gross définit le terme locution comme tout groupe dont les éléments ne sont pas actualisés individuellement.

Pour illustrer ce critère, nous prenons l'exemple de la locution verbale suivante : « *Pierre a pris une veste.* » Gross nous fait constater que la lecture

[1] Ibid, p. 13.
[2] Maria Svensson explique que « L'adjectif qualificatif trémière ne peut modifie que rose, il n'y a pas de forme masculine. »
[3] SVENSSON, Maria Helena, *Critères de figement et conditions nécessaires et suffisantes, Romanesk Forum*, n° 16, 2002, p. 778.

compositionnelle de cette locution permet l'actualisation, par changement de détermination du complément *veste*. Ainsi, il est acceptable de dire *Paul a pris (une, sa, cette, ta, la) veste*. En effet, cette expression est correcte si seulement on la considère dans son sens transparent ou compositionnel, c'est-à-dire dans le sens de (*Il a pris un vêtement*). Au contraire, la lecture de cette expression, comme une expression figée qui veut dire (*être battu aux élections*), fait que le substantif *veste* ne peut pas être actualisé de n'importe quelle façon puisque il ne réfère à aucun vêtement. Et de même, dans la locution adjectivale *à la mode* le substantif *mode* ne peut recevoir aucune détermination autre que le générique *la*.

5. Les études de Maria Svensson

Svensson est de l'avis que nous devons étudier le phénomène de figement à partir de critères et non pas en analysant les catégories et la typologie déjà proposées. Pour ce faire, elle a entrepris une analyse bien détaillée des critères suivants :

-mémorisation.
-contexte unique.
-non-compositionnalité .
-Syntaxe marquée.
-blocage lexical.
-blocage grammatical.

Ces critères rendront compte de la plupart des critères et propriétés proposées par les différents chercheurs en phraséologie, mais dans une terminologie propre à Svensson . Pour situer ces termes dans la terminologie abondante et diverse que connaît ce domaine ,elle présente, à travers le tableau suivant les différentes appellations qui correspondent à chacun de ces termes et adoptées par d'autres auteurs :

Termes adoptés par Svensson	**Termes employés par d'autres chercheurs**	**Nom :**
Mémorisation	*Etymologie*	G .Gross (1996 : 21)
	Préfabriqués ,séquences préformées	Gülich Crafft (1997 : 243,244)
Contexte unique	*Archaïsme ; déficiences lexicales*	Gülich, Crafft (1997 :243,244)

	Eléments archaïques de nature lexicale	Schapira (1999 : 243)
Non-compositionnalité	*Opacité sémantique .*	G .Gross (1996 :10)
	Déficience lexicale et sémantique	Gülich Crafft (1997 :243)
	Figurative meaning	Hudson (1998 :9)
	Restrictions sélectionnelles, valeur intentionnelle, valeur non référentielle.	Martin (1997 : 292, 293)
	Non-compositonality	Moon (1998 :8)
	Conventionality, figuration.	Nunberg et al (1994 : 492)
	Séquence dite opaque	Schapira (1999 : 11)
Syntaxe marquée	*Non-actualistion des éléments*	G .Gross (1996 :13)
	Déficiences syntaxiques, anomalies	Gülich Crafft (1997 :243,266)
	Anomlous syntax or usage	Hudson (1998 :9)
	Eléments archaïques de nature morphologique,éléments archaïques de nature syntaxique,constructions elliptiques.	Schapira (1999 :10)
5-Blocage lexicale	*degré de figement,blocage des paradigmes synonymiques, défigement* *. unexpected collocational restrictions, unexpected syntactic constraints*	G. Gross (1996 :16,17,19) Hudson (1998 :9)
	restrictions sélectionnelles	Martin (1997: 16)
	l'impossibilité de remplacer l'un ou l'autre des mots du groupe,	Schapira (1999 :09)
6-Blocage grammatical	*blocage des propriétés transformationnelles, non-insertion* *-absence de libre actualisation des éléments composant*	G. Gross (1996 :12)

	-impossibilité de changer l'ordre des mots dans la séquence figée, -la suspension de la variation en nombre des composantes, -le segment figé n'admet pas la manipulations transformationnelles -le segment figé ne permet pas l'extraction d'un des composants pour la relativisation, la tropicalisation, la voix passive ou la mise en vedette au moyen de la corrélation ***c'est...que***	Schapira (1999:9)
	unexpected syntaxique constraints	Hudson (1998:8)
	Restriction de la variations ou de transformations	Gülich, Krafft (1997:243)
	Fixedness Variation	Moon (1998: 120-150)
	Inflexibility	Nunberg et al (1994: 492)

Tableau.II.4.Les différents termes employés pour décrire les critères du figement[1]

Nous constatons qu'une grande abondance terminologique règne dans le domaine des critères du figement mais ces termes reçoivent toujours des acceptions plus au moins convergentes. En faites il s'agit des principes que Maria Svensson résume dans les éléments suivants :

- mémorisation (Le rôle de la mémorisation pour les expressions figées) ;
- contexte unique. (Le rôle des mots utilisés uniquement dans les expressions figées) ;
- non-compositionnalité. (La contribution au sens de l'expression par chaque mot qui y figure) ;
- syntaxe marquée. (L'importance des constructions syntaxiques rares) ;
- blocage lexical. (L'impossibilité d'effectuer des commutations) ;
- blocage grammatical. (L'impossibilité de faire des changements syntaxiques).

[1]. SEVNSSON, Maria Helena, *Le critères de figement .L' indentifications des expressions figées en français contemporain,* Print &Media, Umeá, 2004, p.43.

6. La mémorisation :

Nous avons remarqué que l'élément important et originale dans l'étude de Maria Svensson et qui n'est évoqué par les études déjà citées au cours de cette recherche est celui de la mémorisation. De ce fait nous consacrons cette section uniquement à ce critère car les principes des autres éléments sont exprimés par les critères proposés par Gaston Gross et Hudson

Depuis 1997 Grunig a évoqué l' idée que les locution représente un phénomène à fondement psycholinguistique. En effet il leur attribue statut mémoriel :

> *N'importe quelle phrase ou syntagme peut acquérir le statut de titre, ou de phrase historique, ou de rituel – à peu de choses près – même de proverbe, à condition d'avoir un statut social solidaire d'une inscription mémorielle […] ou d'avoir connu un taux de répétition ou notoriété dans une circulation langagière qui les ait transformés en inscriptions mémorielles.*[1]

Grunig insiste aussi sur le figement ou l'immobilité, lequel, il associe à l'inscription mémorielle ce qui fait de la mémorisation un aspect très important du figement .Pour prouver l'existence de ce phénomène, de nombreux études psycholinguistique ont été faites ,notamment sur la compréhension des idiomes ,autrement dit comment accéder au sens que le locuteur a voulu transmettre.

Le premier groupe des psycholinguistes ont confirmé l'hypothèse que les idiomes soient stockés dans notre mémoire et il proposent les modèles suivant :

1) le modèle de la liste mentale d'idiomes, proposé par Bobrow et Bell en 1973,

2) le modèle de la représentation lexicale de Swinney et Cutler, qui date de 1979,

3) le modèle d'accès direct, qui se concentre surtout sur l'interprétation des idiomes. (Gibbs, 1980)[2]

Bobrow et Bell estiment qu'il y a deux moyens séparés de comprendre des phrases : un processus pour interpréter une suite littérale, et un autre pour les expressions idiomatiques. Lorsqu'une interprétation littérale échoue, c'est la liste d'idiomes qui permet à un interlocuteur d'obtenir le sens voulu.

Selon Swinney et Cutler il y aurait également une liste d'idiomes, mais ceux-ci seraient représentés en mémoire de la même manière que les mots. Un moyen spécial pour les

[1] Ibid, p. 45.
[2] Ibid, p. 46

interpréter ne serait pas nécessaire et les deux sens des mots impliqués – littéral et idiomatique – seraient dans ce cas présents en même temps.

Selon Gibbs, finalement, les expressions ambiguës sont interprétées premièrement comme idiomatiques. Mais même s'il constate que l'interprétation idiomatique précède l'interprétation littérale, il en arrive à la conclusion que cela est surtout vrai pour des idiomes

très courants ou familiers. Il est à remarquer que c'est la conventionalité d'une phrase qui décide de la difficulté de compréhension. Dans certains cas, un usage littéral est plus conventionnel, tandis que dans d'autres c'est l'usage idiomatique qui est le plus courant, voire

Conventionnel, ce qui a évidemment une influence sur l'interprétation. Quant aux critiques

qui ont été soulevées contre ces modèles, on notera que ce n'est pas l'hypothèse que les idiomes (dans ces cas-là) seraient mémorisés qui est mise en doute, mais plutôt les procédés

d'interprétation ou d'accès.

Un autre chercheur qui s'appelle Dwight Bolingre s'intéresse, dans un article intitulé *« Meaning and memory* » à la relation entre la mémoire et la production des unités de langue . A ce propos il conclut qu'il y a, en fait, deux espèces de langue : d'un coté la langue « automatique » et de l'autre la langue « propositionnelle »

George Misri [1] a étudié le phénomène de la mémorisation des formules figées en procédant à un examens des reconnaissance inspiré du jeu télévisé *la roue de la fortune* .

Il s'agit de remplacer les mots construits avec « schtroumpf » par d'autres mots. Il a présenté à ses informateurs les phrases suivantes, sans autre contexte :

1. On va schtroumpfer au pont sur la rivière Schtroumpf, aujourd'hui !
2. Schtroumpfer ! Toujours schtroumpfer ! ! J'en ai plein le schtroumpf, moi !
3. Je vais me schtroumpfer dans un coin et piquer un petit schtroumpf !
4. ...espèce de tire-au-schtroumpf !
5. Qu'on ne me schtroumpf sous aucun prétexte !
6. Je ne vous cacherai pas le danger que vous schtroumpferez !

[1] MISRI, Georges, *Le figement linguistique en français contemporain*, thèse de doctorat,Université René Descartes (Paris V) 1987, p. 8-14.

7. Schtroumpfons à la courte schtroumpf !

8. C'est schtroumpfé !

9. Il nous faut la [mouche] « Bzz » schtroumpf que schtroumpf !

10. Schtroumpf-qui-peut !

Misri a rendu compte que les informateurs remplacent la construction schtroumpfs avec les mêmes mots dans la plupart des phrases et parfois avec une variante ce qui donne les constructions suivantes :

2. J'en ai plein le *dos/cul*, moi !

3. [...] et piquer un petit *somme/roupillon* !

4. ... espèce de tire-au-*flanc/-cul* !

5. Qu'on ne me *dérange* sous aucun prétexte !

6. [Je ne vous cacherai pas] le danger que vous *courrez* !

7. *Tirons* à la courte *paille* !

9. [Il nous faut la « Bzz »] *coûte* que *coûte* !

10. *Sauve*-qui-peut !

Misri explique cette unanimité par le fait que ces phrases sont mémorisées par locuteurs . En revanche, pour la phrase numéro 1 (on va schtroumfer au pont sur la rivière schrtoumpf aujourd'hui) les premières parties des phrases 2 et 3 (*Schtroumpfer ! Toujours schtroumpfer !* et *Je vais me schtroumpfer dans un coin*) et la phrase 8 (*C'est schtroumpfé !*), les sujets n'ont réussi de à les complété puisque elle ne sont mémorisé et elles ne constituent pas, donc des expressions figées .Pour pouvoir compléter ces phrases, il nous faut contexte suffisant qui pourrait être visuel comme le cas de la bande dessinée que Misri a testé par la suite.

Grunig arrive à la conclusion que *:« Une locution serait un syntagme complexe inscrit durablement en mémoire et, inversement, tout syntagme complexe ainsi mémoriellement inscrit serait une locution »*[1] alors que Maria Svensson s'interrogent : « *toutes les suites de mots mémorisées, sont-elles des expressions figées ? »*[2]. Cette interrogation est ,en faite pertinente car , les phrases 5 et 6 que l'expérience de Misri a démontré qu' elles sont mémorisées conventionnellement par les sujet , ne pourraient pas être rangées comme des expressions figées et non plus comme des séquences libres .

[1] GRUNIG, Blanche-Noëlle , « La locution comme défi aux théories linguistiques : une solution d'ordre mémoriel ? », in ; Martins-Baltar, Michel , La locution entre langue et usages, ENS Éditions,Fontenay Saint-Cloud.,1997, pp. 225-240.

[2] Svensson, Maria, Op.cit, p. 48

Ce genre de phrases, qui se retrouvent quelque part entre la langue librement engendrée et les expressions figées, sont plus pertinentes lorsque on étudie les *collocations.*

Nous concluons que le critère de mémorisation , évoqué par Maria Svesson permet d'identifier toutes les expressions figées mais il pose un problème ,comme le signale George Misri car li nous amène à identifier, aussi des exemples qui ne sont pas normalement classées comme des expressions figées . Cela nous fait d'avis que la mémorisation constitue une propriété générale du phénomène du figement.

III. Approches linguistiques et enjeux informatiques des expressions figées

1. Introduction

Les expressions figées sont, longtemps restées bien mal inventoriées et étudiées et ce n'est que récemment que le LADL[1] a entrepris une vaste étude de ce secteur du lexique, étude qui renouvelle considérablement la conception qu'on s'en faisait en dépit des obstacles que constituent les variations et le défigement pour ces traitements automatisés.

L'étude proprement linguistique des expressions figées est restée longtemps fragmentaire, en revanche, la curiosité est toujours vive chez n'importe quel locuteur, pour cet aspect de la langue. Le nombre d'ouvrages sur ce point - et pour un public non spécialisé - est important. En voici quelques exemples. la série des *idiomatics* (au Seuil) qui joue du contraste entre la traduction littérale d'une expression figée et son équivalent effectif. En effet "A vue de nez" donne mot à mot "at sight of nose", alors que le correspondant anglais est "at a rough estimate". A cet intérêt pour la sémantique particulière des locutions s'ajoute souvent une fascination pour l'"étymologie" réelle ou supposée de ces expressions. C'est le ressort par exemple du livre de Claude Duneton, *La puce à l'oreille: anthologie des expressions populaires avec leur origine.*

2. L'apport de la lexicographie

2.1 Distinctions entre locution et expression

La lexicographie s'intéresse aussi de ce phénomène, surtout les travaux d'Alain Rey (directeur de la rédaction du *Petit Robert*) qui répondent avec la plus grande rigueur à cette double interrogation à la fois sémantique et étymologique: (Rey 77), (Rey 79), (Rey 86). Il est intéressant à cet égard de noter comment Rey délimite le champ de la phraséologie. Il centre son attention sur la distinction entre les expressions et les locutions qu'il définit comme suit:

> *« [Locution] unité fonctionnelle plus longue que le mot graphique, appartenant au code de la langue (devant être apprise) en tant que forme stable et soumise aux règles syntaxiques de manière à assumer la fonction d'intégrant au sens de Benveniste. C'est pourquoi on peut parler de locutions adverbiales* ou prépositives, *alors que ces mots grammaticaux complexes ne*

[1] LADL : *le Laboratoire d'Automatique Documentaire et Linguistique* du CNRS créé par Maurice Grosse en 1967 et installé d'abord dans le 19e arrondissement de Paris, puis, en 1972, à l'Université de Paris 7 (Jussieu). Avec son équipe, Maurice Grosse lance un vaste programme de description systématique des propriétés syntaxiques de tous les éléments du lexique français en se servant des méthodes de TAL.

seraient jamais appelés des expressions. L'expression est cette même réalité considérée comme un manière d'exprimer quelque chose; elle implique une rhétorique et une stylistique; elle suppose le plus souvent recours à une figur, métaphore, métonymie, etc. »[1]

Nous constatons que Rey exclut de son champ d'investigation les expressions phrases (proverbes...) dans la mesure où elles ne peuvent être définies fonctionnellement comme intégrants d'une unité supérieure à la phrase, unité dont le statut est inconnu, ainsi que les clichés, compris comme suites non codées, modifiables et simplement fréquentes (exemple: un front puissant, un linguiste distingué) . Les *mots complexes* que sont selon lui les locutions fonctionnelles, et les *composés lexicalisés* comme "point de vue" qui sont des syntagmes figés ne reposant pas sur un transfert métaphorique.

2.2. La première conception de la phraséologie

Alain Rey pense que l'intérêt de l'étude des locutions est alors double. Leur fréquence relative et leurs caractéristiques qui peuvent servir à caractériser et à contraster des textes [2]. En se reposant sur ce point de vue nous avons avancé l'hypothèse que les textes journalistiques pourraient être caractérisés par un certain type ou un aspect de figement .

La conception de la phraséologie était, au début "naïve», dans la mesure où l'on s'intéressait restrictivement de sa version lexicographique. Elle voit dans la locution l'exception tant au plan quantitatif qu'au plan formel (constructions et mots employés, mécanismes sémantiques particuliers). La frontière est tranchée entre locutionalité et syntaxe (ou sémantique) libre. L'attention se centre d'abord sur les expressions verbales *(prendre la poudre d'escampette*) et en second sur les noms composés. Ensuite, le besoin compulsif d'explications (étymologiques) témoigne sans doute de la résistance commune à employer des séquences de mots comme des (blocs) au sens global, sans lien aucun avec celui des constituants. Cette répugnance est à mettre en lien avec la fréquence des remotivations partielles ou totales d'expressions figées.

Jusqu'à la fin des années soixante-dix, les expressions figées n'ont pas vraiment retenu l'attention en tant que telles. Considérées en général comme un aspect marginal de la langue par la linguistique traditionnelle, elles ont servi d'arguments dans les débats théoriques entre la grammaire générative et transformationnelle et ses critiques, et au

[1] Rey, Alain, *dictionnaire des expressions et de locutions*, Le Robert, Paris, 1979, p. VI 3.
[2] Ibid, p. XII- XIII

sein de cette nébuleuse même, à divers stades de son évolution conceptuelle. Il en résulte un traitement très fragmentaire et anarchique dans les usuels, tant dictionnaires que grammaires

3. Approches des expressions figées

3.1. L'approche linguistique traditionnelle

Dès 1909, Bally balise le terrain de la phraséologie. Il étudie la place des expressions figées dans la langue maternelle, en lien avec l'apprentissage des langues secondes:

> *« Dans la langue maternelle, l'assimilation des faits de langage se fait surtout par les associations et les groupements dans lesquels l'esprit fait entrer les mots. Ces groupements peuvent être passagers, mais, à force d'être répétés, ils arrivent à recevoir un caractère usuel et à former même des unités indissolubles. Il faut "penser" ces goupements comme le fait le sujet parlant sa langue maternelle. Entre les cas extrêmes (groupements passagers et unités indécomposables) se placent des groupes intermédiaires appelés séries phraséologiques (par exemple les séries d'intensité et les périphrases verbales).»*[1]

Donc, Bally parle en premier temps des associations des groupement ayant un caractère usuel, à force d'être répétés .Il les appelle des séries phraséologiques et il souligne leur importance pour l'assimilation d'une langue seconde comme le fait le sujet parlant dans sa langue maternelle .Il s aperçoit, par ailleurs qu'il s'agit d'un continuum entre séquences entièrement figées et suites simplement "durcies". On opposera par exemple:

> *sans contredit ;*sans (un + ce + le) contredit,*sans contredits, *sans contredit important.*
>
> et
>
> *à (la + E) condition (de + que), à la condition impérative (de + que)*

La première séquence présente un figement totale alors que la deuxième tolère quelques insertions . Dans la traduction anglo-saxonne on parle la partition entre *canned phrases* (expressions toutes faites) et flexible idiomes .

[1] BALLY, Chales , Op.cit, p .66.

La tradition linguistique s'appuie sur un certain nombre de critères formels pour déterminer si une suite de mots est figée:

1- Processus d'intégration linguistique des composés

a- au niveau phonique: la non-prononciation du f dans *chef d'oeuvre*, ou l'absence de liaison entre *boîtes* et à dans *des boîtes à ordures* signalent la constitution d'un groupe complexe.

b- à l'écrit, l'alternance avec une version comportant des traits d'union *(c'est à dire / c'est-à-dire*) ou l'existence d'un sigle (*RIB / relevé d'identité bancaire*) ou d'une abréviation (*et cætera / etc.*) constituent d'autres indices de figement.

2- contraintes sur les modifieurs possibles:

pour l'expression au *passage Xavier a réagi*, l'ajout d'un modifieur **au sombre passage, Xavier a réagi donne lieu* à une séquence inacceptable.

3- impossibilité de substitution d'un antonyme ou d'un synonyme:

à propos, Xavier a appelé

**à sujet, Xavier a appelé*

4-- impossibilité de reprise partielle:

le chemin de fer est en crise

**le chemin est en crise*

le général de division se porte bien

le général se porte bien

Mais, comme nous avons vu dans les chapitres précédents, les partitions effectuées entre séquences libres et figées s'appuient essentiellement sur la démotivation éventuelle de tout ou partie des constituants, ce que l'on appellera une sémantique non compositionnelle: le sens de la suite de mots ne se déduit pas de la simple combinaison de ses termes, en fonction de la structure syntaxique présente. Dans ce domaine, un continuum reste observable, des groupes dont tous les constituants sont sémantiquement absents *.de plein fouet.*. à ceux dont tous les constituants sont sémantiquement présents *sans arrêt.,* via ceux qui associent des termes vides à des éléments sémantiquement présents .en désespoir de cause..

3.2.1. L'apport de la sémantique

La sémantique n'étant pas la branche de la linguistique la mieux formalisée, ni celle dans laquelle les points d'accord dominent, rien d'étonnant à ce que cette approche sémantique prête le flanc à la critique. Ainsi la paraphrase par un mot unique *.chien assis.* (*lucarne*), souvent invoquée, ne témoigne pas forcément de l'unité

sémantique de la séquence. Elle n'est d'ailleurs pas toujours praticable, même pour des séquences simples, fréquentes, et dont le sens se calcule aisément: c'est le cas d'*avoir faim / froid* D'après Benveniste, le signifié a un caractère unique et constant. En effet, il est simple à établir dans le domaine concret (*.chaise électrique / roulante / percée.)* mais il l'est nettement moins quand il s'agit de concepts. Ainsi en va-t-il pour les expressions figées , plus spécifiquement, dans des langues de spécialité (terminologie d'un domaine), une connaissance du domaine et des concepts qu'il met en oeuvre est indispensable pour établir si une suite de mots donnée correspond ou non à une entité abstraite.

4. les expressions figées et la grammaire générative et transformationnelle

l'analyse des expressions figées dans l'ombre de la grammaire générative et transformationnelle, ou dans la critique de cette école, n'a jamais constitué un domaine à part entière. La typologie des verbes figés en fonction du nombre de transformations observées dans une séquence a simplement pour but d'intégrer les *idiomes* dans une sémantique compositionnelle; elle ne repose pas sur une étude extensive et intensive de tels verbes. (Bresnan 82) entend seulement montrer que l'on peut rendre compte du comportement des verbes figés face à la passivation en faisant l'économie des transformations. (Ruwet 83) étudie les expressions verbales de la forme [V GN] (*tirer le diable par la queue, faire feu...)* pour contester sur ce point, et peut-être même de façon plus globale, les thèses de Chomsky. Malgré un nombre relativement important de verbes complexes examinés, Ruwet n'échappe pas à une vassalisation par la théorie des données empiriques. En effet, il donne une liste importante de propriétés, c'est-à-dire de constructions, à examiner pour cette structure [V GN] et il plaide ainsi indirectement pour une approche syntaxique du figement, libérée le plus possible de toute dimension sémantique.

5. Les mots composés dans les dictionnaires et les grammaire

Le trait d'union est l'objet d'une attention particulière dans la mesure où, sa présence entraîne pour le mot une entrée autonome, à sa place alphabétique; son absence entraîne parfois l'omission pure et simple du mot, ou le relègue dans une rubrique annexée à l'article d'un de ses composants. On comparera simplement par exemple *c'est-à-dire* et *c'est pourquoi*.(Catach 81) donne un échantillon de la variation du traitement d'un certain nombre de mots composés sur ce point dans trois dictionnaires généraux, *le Petit Larousse, le Littré et le Robert.*

Contrairement à des langues comme l'allemand qui usent dans ce cas de la soudure, le français pour ce faire ne recourt guère à des marques particulières (on trouve cependant des alternances trait d'union - soudure: *delta-plane* / *deltaplane*). Il faudra donc distinguer pour certains caractères les cas où ils sont effectivement des séparateurs et ceux où ils ne le sont pas. C'est le cas de l'apostrophe: *l'héroïne* / *c'est-à-dire* / *c'est*, du trait d'union: *donne-la / tiers-monde*, et surtout du blanc.

Catach souligne aussi l'arbitraire dans les mots composés retenus par les usuels:

> *« le mot composé est un syntagme. Un dictionnaire enregistre en principe les mots, non les syntagmes. Toute unité sémique qui ne sera pas en même temps graphique risque fort d'être arbitrairement rejetée aussi bien que retenue. »(…)*
>
> *« On ne trouvera nulle trace à l'ordre alphabétique de locutions et syntagmes aussi courants que" bon marché" adjectif, "collet monté" adjectif, "salle à manger", "arc de triomphe", ou "trait d'union" (qui, comme mot composé, n'a pas de trait d'union). Quels sont, à l'heure actuelle, les dictionnaires qui accordent à(chemin de fer, hôtel de ville, et pomme de terre) leur place à l'ordre alphabétique ? »*[1]

Dans le même ouvrage Catach affirme que l'entrée en mémoire du potentiel lexical de la langue risque d'être extrêmement partielle (bon nombre de néologismes compositionnels n'étant pas reconnus comme tels).

La position manifestée par les grammaires traditionnelles est la stricte conséquence du peu d'intérêt théorique et empirique à l'égard du figement jusque récemment. La question n'y figure pratiquement pas, à part les considérations sur l'écriture des mots composés, où la rationalité affichée et les explications fournies sont quelque peu battues en brèche par les contradictions et les incohérences que montrent les ouvrages de référence sur ce point. (Mathieu-Colas 87, p 3) indique par exemple que le *Petit Robert* (édition de 1984) parle de "serpent à sonnettes" aux entrées "*serpent" et "sonnette" et de "serpent à sonnette" à l'entrée "crotale".*[2]

[1] CATACH, Nina, *Orthographe et lexicographie*, Nathan, 1981, p. 7,17.
[2] *Enjeux linguistiques et informatiques des expressions figées* , www.limsi.fr/Individu/habert/Publications/Fichiers/habert91b/BH_C1.html .

6. les expressions figées et le Traitement automatique des langues(TAL)

6.1. Les travaux de LADL

Maurice Grosse, ingénieur de formation qui s'était intéressé aux travaux de .Noam Chomsky et M P. Schützenberger , au cours de ses deux séjours aux Etats-Unis, avait rencontré de nombreux linguistes et avait collaboré avec Z. Harris à l'Université de Pennsylvanie (1964) .En 1967, il est entré au Laboratoire de calcul Blaise-Pascal du CNRS, travaillé avec M. P. Schützenberger et publié, avec André Lentin, *Notions sur les grammaires formelles*. En 1968, il a créé le Laboratoire d'Automatique Documentaire et Linguistique (LADL, laboratoire du CNRS) et publié *Syntaxe du verbe* (Larousse, 1968), premier tome de sa série *Grammaire transformationnelle du français*.

Le laboratoire de recherche a été installé, d'abord dans le 19e arrondissement de Paris, puis, en 1972, à l'Université de Paris 7 (Jussieu). Avec son équipe, Maurice Grosse lance un vaste programme de description systématique des propriétés syntaxiques de tous les éléments du lexique français. Il s'appuie pour cela sur les théories distributionnelles et transformationnelles de Z. Harris, avec lequel il a toujours gardé une grande connivence intellectuelle et scientifique.

Le LADL a entrepris une description systématique des expressions figées du français. Cette mise à plat s'inscrit dans la lignée des travaux précédents sur les formes simples, dont il faut rappeler les fondements théoriques, méthodologiques et les résultats, notamment ceux de Z. Harris :

> *« L'hypothèse fondamentale de la théorie transformationnelle de Z. S. Harris distingue les phrases élémentaires ou noyaux comme unités de base de la composition syntaxique. Les phrases élémentaires sont sémantiquement invariantes par transformation. La systématisation de cette hypothèse dans la théorie du lexique-grammaire conduit à considérer la phrase élémentaire comme unité sémantique de base et non pas le mot. Dès lors, deux séries de vérifications empiriques de cette hypothèse sont nécessaires:*
>
> *1) on doit montrer qu'aucun mot de la langue n'a d'autonomie syntactico-sémantique, autrement dit que tout mot entre dans une phrase élémentaire caractéristique;*
>
> *2) on doit vérifier que toute phrase complexe s'analyse en termes de phrases élémentaires. »*[1]

[1] GROSS, Maurice, , « Les limites de la phrase figée », Langages, Larousse, Paris. n°90, 1988, p47.M

Cette étude du lexique se base sur des phrases élémentaires construites pour être analysées. Les énoncés déjà existants qu'ils soient oraux ou écrits ne sont pas utilisés car ils sont généralement trop longs et sources d'ambiguïtés multiples. Les phrases sont tout d'abord soumises à un jugement d'acceptabilité pour déterminer si la phrase élémentaire construite est grammaticale ou ne l'est pas. Les mots qui composent ces phrases sont ensuite analysés selon leur contexte et leurs cooccurrences. L'étude d'un mot donné aboutit à l'émergence d'un certain nombre de propriétés. Des professionnels de la linguistique ont, ensuite pour tâche de valider les propriétés définies pour chaque mot.

Les travaux du LADL ont, naturellement évolué vers une étude extensive des expressions figées du français, naturellement, car une étude de chaque mot du lexique ne pouvait que laisser présager une telle démarche. En effet, les expressions figées sont constituées de mots simples qui ont donc, selon les contextes tantôt un emploi libre tantôt un emploi figé. Il était donc « logique » que les expressions figées prennent davantage

d'importance dans l'approche du l LADL. Ce groupe de recherche a , en fait entrepris de recenser toutes les expressions figées, ce qui a permis de mesurer, au sens propre le poids de ces expressions dont le nombre est nettement supérieur à celui des formes libres. L'ensemble des études menées par le LADL aura donc eu pour conséquence de faire du figement un objet linguistique autonome.

Les divers résultats produits par ces recherches sont reproduits sous forme de tables. Ces différentes tables représentent le lexique-grammaire élaboré au LADL. Les tables qui composent ce lexique-grammaire regroupent tous les éléments du lexique.En effet, chacune de ces tables contient un ensemble de propriétés qui s'établissent en colonne. En vis-à-vis de ces colonnes, un codage avec un signe positif « + » ou négatif « - » permet de préciser si l'élément du lexique figurant dans la table peut être défini ou non par cette(s) propriété(s). Ces tables, éditées sous formes des fichiers Excel au format « .xls » constituent actuellement des ressources électroniques pour les recherches automatiques des expressions figées.

Nous pouvons donc constater, aussi que le LADL a une démarche morphologique quant au traitement des corpus. En effet, le filtrage des séquences

complexes correspond à la reconnaissance des propriétés syntaxiques définies par les tables.[1]

6.2. l'élaboration du lexique-grammaire du français

Maurice Grosse a constater, par ailleurs, que N. Chomsky avait pertinemment évoqué le problème des contraintes lexicales dans les opérations transformationnelles, en particulier dans *Aspects of the Theory of Syntax* (1965), mais il l'avait laissé de côté une grande partie de la structure lexicale :

> *« Une grande partie de la structure lexicale n'est, en fait qu'une classification entraînée par le système des règles phonologiques et syntaxiques. Postal a, de plus, suggéré qu'il devrait y avoir une analyse générale des éléments lexicaux du point de vue de chaque règle R, les divisant entre ceux qui doivent, ceux qui peuvent, et ceux qui ne peuvent pas être soumis à la règle R [...]. Je mentionne ces possibilités afin seulement d'indiquer qu'il reste plusieurs façons relativement inexplorées de traiter les problèmes qui se posent lorsque l'on considère avec sérieux la structure du lexique. [...] Pour le moment, on peut à peine dépasser le simple arrangement classificatoire des données. Quant à savoir si ces limitations sont intrinsèques ou si une analyse plus profonde peut parvenir à débrouiller certaines de ces difficultés, cela reste en suspens. »* [2]

Les travaux du LADL se sont éloignés progressivement, puis de manière tranchée, des perspectives ouvertes par Chomsky. Pour le LADL en effet, Chomsky et ses épigones, s'ils ont mis en évidence de nombreux phénomènes syntaxiques, n'ont jamais étudié systématiquement le comportement de ces propriétés en fonction des éléments du lexique. L'accent a été mis sur la formalisation des phénomènes. Les généralisations se trouvent alors fondées sur un nombre nettement trop restreint d'exemples.

Ce problème en suspens, M. Gross décide de s'y attaquer, seul d'abord, puis avec son équipe du LADL (CNRS). Il écrit dans *Méthodes en syntaxe* :

> *« Les études transformationnelles ne portent que sur un petit nombre d'exemples. Elles ont dégagé un grand nombre de phénomènes nouveaux, mais*

[1] LE ROI, Marie-Véronique , Op.cit .41.
[2] CHOMSKY, Noam , *Aspects of the Theory of Syntax,* Trad. De J-C. Milner.1971.

elles ne permettent pas d'évaluer l'étendue de ces phénomènes pour une langue donnée."[...]

« Après une période où des succès ont pu laisser croire que l'emploi de transformations dans les descriptions allait régulariser considérablement ces dernières, il est devenu clair que les nouvelles règles continuaient à comporter des "exceptions" en nombre sensible. Il est donc devenu crucial de vérifier ces théories en entreprenant la description d'une langue au moins [...]. »[1]

En revanche, le LADL a choisi de reléguer au second plan la formulation de théories générales, au profit de l'accumulation de données sur le fonctionnement concret des mots du lexique. Cette accumulation ouvre sans doute la voie à des généralisations, mais celles-ci ne doivent pas anticiper sur le travail de classement. Le modèle du français est alors limité et peu formalisé. Et l'étude du français entend être extensive (étudier la majeure partie du lexique commun) et intensive (étudier pour chaque item lexical le maximum de propriétés connues).

Sur 20 ans, entre une dizaine et une vingtaine de linguistes ont ainsi défriché les propriétés des mots simples, partie du discours par partie du discours.

Il s'agit alors pour (tous)les mots du lexique commun d'en étudier systématiquement les propriétés syntaxiques, c'est-à-dire d'examiner dans quelles constructions entre chaque mot. C'est ce qu'on appellera un lexique-grammaire. Cette théorie écarte l'approche sémantique jugée trop subjective et variante d'un linguiste à l'autre., dans la mesure où les jugements sur ce plan n'apportent pas assez de garanties de reproductibilité (d'un linguiste à l'autre ou même au fil du temps) pour être vraiment utiles. Par ailleurs, quand un même mot peut rentrer simultanément dans divers ensembles de propriétés, c'est certainement couplé à des distinguos sémantiques. Ainsi, on peut isoler deux "*au passage"*

au passage de Pierre, Xavier a tiqué

à son passage, Xavier a tiqué

s'opposera à:

au passage, Max signale qu'il arrive demain

qui n'a pas d'équivalent:

**au passage de X, Max signale qu'il arrive demain*

[1] GROSS,Maurice, *Méthodes en syntaxe*,Hermann, 1975.

L'étude des transformations, c'est-à-dire des structures apparentées pour une construction de départ donnée, permet alors de découvrir les contraintes propres à chaque élément du lexique (ou éventuellement, comme pour "au passage", de dégrouper un mot en fonction des divers ensembles de propriétés dans lesquels il rentre). Par exemple, pour les adverbes complexes de la classe PCDN de la forme [Prep Det X Prep] les propriétés suivantes, entre autres, seront examinées:
Equivalence de l'adverbe avec un régime et d'une forme avec déterminant possessif sans régime:

Sur ses conseils, Max est revenu
Sur ses conseilles, Max est revenu
Au lieu de Pierre, Max est revenu
**A son lieu, Max est reve*nu
Equivalence avec une forme comportant un démonstratif :
A propos de bottes d'oignons, Max a appelé
A ce propos, Max a appelé A l'aide de Pierre, Max a réussi
Au lieu de pierre, Max est revenu.
**A cette aide, Max a réussi*

De la même manière, les distributions d'un mot, c'est-à-dire les formes avec lesquelles il co-occurre, seront étudiées . Il peut être nécessaire parfois de lister les formes acceptées, les restrictions présentant trop de particularités pour une définition en intension. Gross Maurice donne l'exemple de la construction adverbiale: *en toute N*, qui admet comme *N* un nom de qualité, mais avec des interdictions imprévisibles: *en toute simplicité* / *en toute intelligence. Des noms classifieurs, qui abrègent toute une classe, peuvent aussi être utilisés. Par exemple, le nom classifieur *Ntemps*, qui rassemble des mots comme: *jour, journée, mois, semaine, année*..., peut servir à décrire les adverbes qui entrent dans le schéma: LE Ntemps (dernier + précédent + suivant + d'après), comme *le jour (suivant + précédent), le mois d'après* etc. Enfin, la distribution peut dans certains cas être définie purement en intension, par un ensemble de traits sémantiques, ensemble limité en nombre et en complexité pour qu'il puisse être utilisé de manière cohérente au fil du temps et d'un linguiste à l'autre.

Pour que les propriétés attribuées soient fiables, un certain nombre de précautions et de principes président à leur attribution

En premier lieu, porter un jugement d'acceptabilité sur une phrase suppose que l'on maîtrise le sens des mots mis en relation. Le travail se restreint alors à dessein au français général, éliminant les formes archaïques ou trop techniques.

En second lieu, porter un jugement d'acceptabilité représente pour la linguistique ce qu'est l'expérience pour d'autres sciences, c'est-à-dire la vérification soigneuse, dans des conditions bien spécifiées, des propriétés examinées. Cela revient à travailler sur des phrases élémentaires, construites pour l'occasion, plutôt qu'à recourir à des énoncés effectifs, écrits ou oraux, toujours trop longs et porteurs d'ambiguïtés multiples. C'est donc une tâche de professionnels de la linguistique, qui ne peut être remplacée par le questionnement de locuteurs natifs.

Enfin, la vérification de l'acceptabilité de telle construction par croisement des jugements de linguistes professionnels est cruciale. D'où par exemple la publication des résultats concernant une série de formes sous forme de tables soumises à discussions et à vérification.

6.3 L'approche automatique des expressions figées

L'étude systématique du lexique a conduit à donner aux expressions figées la place qui leur revient, ne serait-ce que sur un plan quantitatif. Il fallait également vérifier si les hypothèses générales sur la phrase élémentaire étaient confirmées ou infirmées par l'examen des figements[1] . L'accent mis sur les contraintes pesant sur les distributions et les transformations des différentes entrées d'un lexique-grammaire ne pouvait que rendre sensible aux *figement* au sein d'un syntagme ou au sein de la phrase. De plus , un intérêt accru pour les *industries de la langue* a certainement contribué à l'étude de phénomènes particulièrement importants dans les domaines techniques (terminologie) et pour la traduction automatique. Ceci a fait l'étude extensive et intensive des expressions figées du français l'un des objectifs majeurs du LADL.

6.3.1. les contraintes particulières du figements

Les données de fréquence, l'impression d'avoir affaire à un cliché, l'explication du sens des expressions figées sont écartées Le figement se caractérise avant tout par les contraintes particulières qui portent sur une séquence de mots, soit dans la combinatoire interne de ses constituants (impossibilité ou au contraire obligation de modifieurs ou de déterminants au sein d'un syntagme nominal: *par exemple*/ **par un exemple , à brève*

[1]GROSS, Maurice, Op.cit, p 7-22

échéance / **à échéance* ...) soit dans la combinatoire de la suite de mots au sein de la phrase entière (**il a répondu en choeur*). A cette option fondamentale, peut s'ajouter le fait que la sémantique de la séquence est manifestement non compositionnelle: *dépérir à petit feu*.

Les divers travaux soulignent la complexité de la tâche, mais surtout l'inanité de critères généraux. Ce qui est dit des noms composés peut être élargi à toutes les parties du discours.

On retrouve sur ce point la démarche qui est celle du LADL sur les formes simples: juxtaposer les études limitées et détaillées sur des secteurs bien déterminés (les phrases figées (Gross. M 88), l'adverbe (Gross. M 84)(Gross. M 90), le nom (Gross. G 86), les connecteurs (Piot 88), telle construction: être Prép X de N2 (Danlos 81), pour couvrir les diverses manifestations du phénomène sans déboucher sur des généralisations prématurées.

(Gross G. 88) donne pour le patron [N Adj] un certain nombre de propriétés à examiner pour déterminer le caractère (plus ou moins) figé ou non figé d'une suite de mots. En voici quelques unes.

- L'adjectif peut-il être placé en position prédicative ?

un livre difficile / *ce livre est difficile*

l'impossibilité de la prédication est souvent fonction de l'emploi métaphorique de l'un ou l'autre élément ou des deux: *un champignon atomique* : **ce champignon est atomique*; *une arme blanche* : **cette arme est blanche*; *une chambre forte : *cette chambre est forte*[1]

- L'adjectif peut-il être nominalisé ?

On opposera:

(*Max + ce témoignage) est fragile*

(Max + ce témoignage) (a + est d) une certaine fragilité

à

Une messe noire

** la noirceur de cette messe*

L'existence de (trous morphologiques) (tout adjectif ne possède pas de correspondant nominal) et des restrictions diverses vient contrarier l'effet de ce critère.

- Le nom ou l'adjectif peut-il se voir substituer des mots de la même série sémantique ?

[1] GROSS, Gaston, *Degré de figement des noms composés*, Langage, n°90, 1988, pp. 64.

*une douche (écossaise + *française + *norvégienne*)

*un garçon manqué | une *fille manquée.*

Gaston Gross affirme que « *Le figement peut ne pas être total. Il se forme alors un début de paradigme »*

*Max est un col (bleu + blanc + *vert)*

*Max est UN (cerveau + tête + *bras) brûlé(e)*

- La variation en nombre est-elle libre ?

les eaux usées de Meudon

**l'eau usée de Meudon*

les forces vives de la nation

**la force vive de la nation*

- Un adverbe peut-il modifier l'adjectif ?

Paul est col blanc

**Paul est col très blanc.*

« des groupes nominaux interprétés intuitivement comme relativement figés peuvent accepter entre le nom et l'adjectif certains éléments comme "dit", "appelé", "prétendu" etc., et des adverbes "prétendûment", "soi-disant"."une grève dite tournante".

*Cette construction ne semble possible que si le figement n'est pas total: *(un + ce) blouson dit noir, *du fer dit blanc, *une bande dite dessinée, *un col dit vert. Dans les groupes nominaux ordinaires, cet emploi du participe "dit" est très peu naturel: "?un meuble dit cassé", "?une page dite raturée". »*[1]

Une expressions métalinguistique de ce type indique d'ailleurs à la fois que le figement n'est pas total, puisque l'expression se laisse partiellement disloquer, et que le locuteur a conscience du lien particulier que les composants entretiennent.

- L'adjectif peut-il être coordonné à un autre adjectif ?

un livre difficile mais intéressant

Paul est un col blanc

/* Paul est un col blanc et bleu

- L'adjectif peut-il être effacé et le groupe repris par le N seul ?

le bras droit de N / *le bras de N

*« Dans un groupe nominal habituel, on ne peut pas effacer le substantif-tête: "un livre difficile", *"un difficile". Quand l'adjectif est interprété*

[1] Ibid, p. 67.

comme classifieur et non comme qualificatif, l'effacement du substantif n'est pas exceptionnel:

L'école communale / la Communale

du vin rouge / du rouge »[1]

- Le nom du groupe N Adj peut-il être repris en position prédicative ?

Dét N Adj est Dét N

**un col blanc est un col*

Gross conclut que :

« les composés totalement figés, ceux qui ne prêtent matière à aucune variation, ne représentent que 10 % des groupes nominaux qui, à un titre ou à un autre, n'ont pas toutes les propriétés habituelles des structures dont ils relèvent. On doit alors considérer la composition comme une échelle de figements dont les valeurs limites ne doivent pas former des entités spécifiques. »[2]

Il s'agit donc d'un nouvelle tentative pour donner un contenu plus précis à la notion intuitive de "degré de figement" et pour corréler liberté syntaxique et compositionalité sémantique.

6.3.2. La reconnaissance des expressions figées

Dans les différents systèmes d'analyse syntaxique automatique existants, les expressions figées sont rarement l'objet d'une attention en tant que telles (la situation est d'ailleurs en train de changer .Elles font plutôt figure de "scories" de l'approche syntaxique et / ou morphologique selon les cas.

« Les systèmes existants ne prennent pas en compte [les expressions figées] d'une façon satisfaisante, car on manque de méthodes générales pour les reconnaître. »[3]

En particulier, n'a pas été réglée la question de savoir si elles doivent être traitées dans la phase lexicale ou dans la phase syntaxique, ou bien si les diverses expressions doivent être réparties entre ces deux étapes .

La prise en compte des mots composés apparaît cependant comme une étape importante pour la conception de certaines interfaces entre l'homme et la machine, et plus

[1] Ibid, p. 68
[2] Ibid, p. 70
[3] Laporte, Eric, "*La reconnaissance des expressions figées lors de l'analyse automatique in Les expressions figées", Langages*, vol. 23, n°90. 1988, p 177.

généralement pour l'analyse automatique du français. On peut attendre en particulier du repérage des expressions figées une simplification sur certains points des grammaires à produire pour l'analyse automatique:

- "nettoyage" d'irrégularités difficiles à intégrer et dues à des traces d'états antérieurs de la langue: "*à vrai dire*", "*sans mot dire", "chemin faisant"...*
- repérage de segments issus de langues étrangères et intégrés tels quels dans les textes: locutions latines *("ad hoc", "mutadis mutandis"* ...), anglaises ("*up to date*" ...)...
- "élagage" d'arbres syntaxiques: la construction comme unités complexes de syntagmes nominaux comme "caisses de régimes de retraite complémentaire" "aplatira" l'arbre de la phrase et le simplifiera. En outre, elle contribuera à lever les incertitudes qui peuvent exister quant au rattachement de tel syntagme.

> *« La détection d'une expression figée sans analyse du contexte apporte des informations servant à l'analyse du contexte. Dans l'exemple*
> (*Paul a parlé à son fils à temps),*
> *l'identification de 'à temps' comme adverbe figé donne accès à ses propriétés syntaxiques. En particulier, il s'agit d'un adverbe de phrase, ce qui permet de rattacher ce groupe prépositionnel directement au verbe 'a parlé', et non au nom 'fils'. Cela implique, pour la suite de l'analyse, que le groupe nominal formé autour du nom 'fils' est réduit à 'à son fils', et que l'objet indirect en 'à' du verbe 'a parlé' ne peut être que 'à son fils', à l'exclusion de 'à temps'. Autrement dit, on dispose d'informations sur les limites du groupe nominal voisin et sur son attachement ».*[1]

7. les méthodologies de la reconnaissance automatique des expressions figées

Le nombre extraordinaire de textes électroniques disponibles du au développement prodigieux des moyens de communication notamment l'Internet a rendu, depuis quelques années le traitement automatique des langues naturelles et ses applications incontournables. La plupart des outils adaptées utilisent des approches statistiques. Cependant, depuis longtemps, les chercheurs connaissent l'intérêt d'intégrer à ces systèmes de vastes bases de données de descriptions linguistiques fines. Dans cette optique, le laboratoire d'Automatique Documentaire et Linguistique puis le

[1] Ibid., p.122.

réseau de laboratoires européens RELEX [1]accumulent depuis les années soixante-dix une large variété de composants linguistiques où le lexique joue un rôle fondamental. Avec l'aide d'une méthodologie claire et rigoureuse et de la technologie à états finis (ROCHE 97 ; MOHRI 97), de larges dictionnaires et grammaires ont été créés et appliqués à des textes avec les logiciel Intex (Silberztein 93, 94) et Unitex ([Paumier 02) et leurs extensions ([Laporte 99 ; Paumier 00). Actuellement, nous assistons à une augmentation spectaculaire du nombre de ressources notamment des grammaires sous la forme de graphes. [2]

7.1. Les différents niveaux d'analyse

Les grammaires sous la forme de graphes constituent un outil qui permet de différents niveaux d'analyse des textes et que nous allons expliquer plus en détaille dans le chapitre suivant . Nous distinguons d'abord deux niveaux selon l'unité minimale utilisée (caractère ou mot). Lorsque l'unité minimale est le caractère, nous pouvons parler de traitement morphologique. Dans ce cas, les graphes utilisés servent à décrire des variantes orthographiques de manière compacte et donc à alimenter les dictionnaires électroniques. Nous nous intéressons maintenant au cas où l'unité minimale est le mot. Les niveaux d'analyse y sont plus nombreux. Tout d'abord, les graphes peuvent être assimilés à des extensions des dictionnaires des mots composés. Par exemple, la description des dates sous la forme d'automates factorise de façon significative un ensemble d'expressions quasiment impossible à traiter sous forme de listes (Maurel 90 ; Baptista 99). Ensuite, il est possible de décrire les contraintes locales autour d'un mot de manière très fine. Ainsi, nous pouvons constituer des classes de mots composés ayant un sens proche, comme le graphe *Station*. Ce dernier graphe permet notamment de distinguer deux entrées lexicales de *station* : (1) *station (E + de ski)* et (2) *station (E + de métro)*. J. Senellart (1998) a construit des grammaires pour des noms d'activités telles que *ministre (E+ de l'intérieur)*. L'étape suivante est de construire des constituants de phrases comme les groupes nominaux ou les groupes verbaux comme le

[1] Le réseau RELEX est un ensemble informel de laboratoires européens travaillant dans les domaines de la linguistique et du traitement automatique des langues naturelles. Les différentes équipes travaillent sur un nombre important de langues comme le français, l'anglais, le portugais, l'allemand, l'espagnol, le norvégien, le coréen, le thaï,… Elles utilisent une méthodologie commune : le lexique-grammaire. Le lexique y occupe une place fondamentale, ce qui se traduit par la construction de bases de données linguistiques à large couverture. Le logiciel INTEX sert de plate-forme linguistique commune pour appliquer ces ressources à des textes réels.

[2] CONSTANT, Mathieu, *vers une construction d'une bibliothèque en-ligne de grammaires linguistiques*, http://ladl.univ-mlv.fr p.2, consulté le 12/2/2008 .

montre M. Salkoff (1973) pour construire une grammaire en chaîne du français. Afin de reconnaître automatiquement des expressions figées dans les corpus, J. Senellart (1999) a élaboré quelques grammaires de groupes nominaux simples comme montré dans le graphe ci-dessous. C. Dominguès (2001) a regardé le comportement de groupes nominaux contenant une coordination. La constitution de grammaires complètes reconnaissant les GN est l'un des futurs enjeux du réseau RELEX.

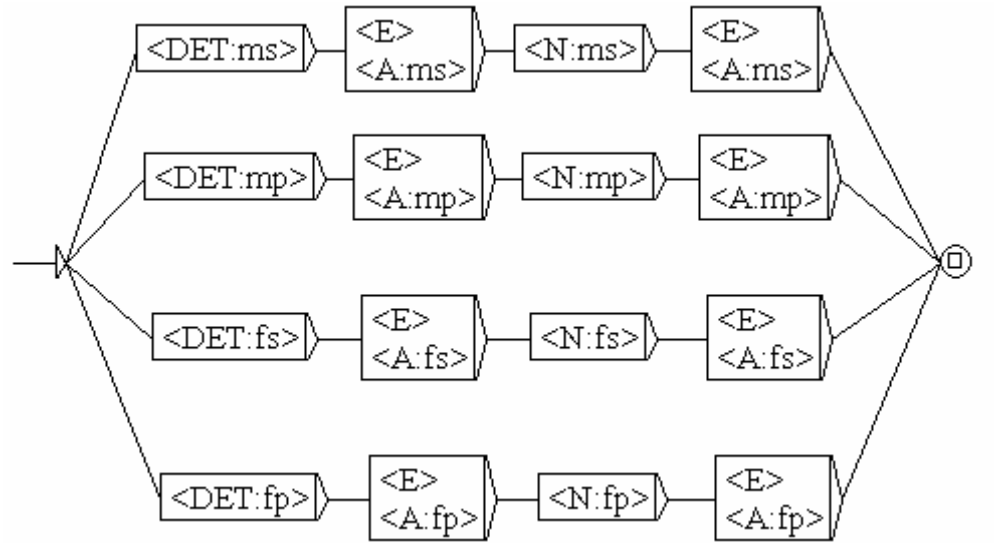


Schéma.III.1.Graphe des GN simple de J. Senellart[1]

A partir de l'étape précédente, il est possible de décrire des phrases simples libres contenant un prédicat (verbe, nom, adjectif) et des arguments comme l'a fait E. Roche (1993, 1999) à l'aide de tables de lexique-grammaires et de transducteurs à états finis. Un travail de grande envergure dans la continuité de cette étude est actuellement mené au sein de l'université de Marne-la-Vallée. J. Senellart (1999), à l'aide de graphes, a décrit les expressions figées du français, à partir des tables de M. Gross (1983) et montré par la même occasion leur présence en grand nombre dans les textes français et plus particulièrement dans les textes journalistiques . Elles en constituent 30% selon Senellart . Ce qui confirme notre hypothèse pour ce genre de texte .

7.2. La « Zone Fixe » des expressions figée

L'étiquetage une opérations qui est d' une grande importance pour le traitement automatique d'un corpus présuppose et pour que tous les mots qui figurent dans ce corpus soient connus et identifiés.. Cette opération consiste donc à fournir des informations morphosyntaxiques sur les mots qui composent les différentes phrases d'un corpus donné. Les informations fournies par l'étiquetage différeront selon les buts

[1] Ibid , p. 7

visés par l'analyse. Un même mot peut aussi bien avoir un emploi libre qu'un emploi figé : l'étiquetage ne prend en compte que les mots simples. Cette étape à elle seule ne permet donc pas de reconnaître les locutions verbales. Eric Laporte (*1988*) introduit la notion de zone fixe pour décrire un mode de reconnaissance automatique des expressions figées. La « zone fixe » d'une expression figée désignerait la partie de l'expression qui admet un nombre de fixe mots simples, même si ces mots sont susceptibles de variations morphologiques ». Dans le cas des séquences verbales, les verbes supports sont exclus de la zone fixe. Dans une expression telle que « *être bon public* », la zone fixe se limitera à « bon public » dans la mesure où le verbe « être » est un verbe support et qu'il peut donc être effacé ou remplacé par une variante aspectuelle.[1]

- *être bon public : Pierre est bon public*
- *Pierre a des amis bons publics. (effacement)*
- *Pierre est devenu bon public. (remplacement par une variante aspectuelle)*

Donc "*être "*est un verbe support.

La reconnaissance de la zone fixe permettrait d'aboutir à la constitution d'une base de données contenant les formes de différentes expressions figées existantes ainsi que leurs

propriétés. Cette méthodologie permettrait de reconnaître automatiquement les expressions

figées dans la mesure où les formes données par le dictionnaire apportent aussi des informations distributionnelles. La zone fixe d'une locution comme *« casser sa pipe » serait* donc décrite de la manière suivante :

N0 Casser Poss pipe

N0 est une variable désignant le groupe nominal Sujet. Le possessif Poss est donc variable..

La zone fixe permet donc de reconnaître des séquences figées même lorsque ces dernières connaissent des variations.

7.3. Les méthodes statistiques et / ou structurelles

L'analyse des expressions figées dans un corpus donné au moyens des outils du Traitement Automatique de la langue pourrait se faire en procédant à deux méthodes principales : une approche statistique ou une approche structurelle. La méthode

[1] LE ROI, Marie-Véronique, Op.cit, p. 41

statistique a la particularité de ne nécessiter qu'un nombre limité de connaissances linguistiques. L'approche structurelle, quant à elle, requiert davantage de connaissances linguistiques. En effet, un outil statistique n'a recours qu'à un lexique de mots fléchis et leurs catégories pour assigner des étiquettes grammaticales aux mots d'un texte.

Dans les études actuelles , nous avons généralement recourt un analyseur statistique puis à un étiqueteur structurel. En effet nous allions les deux méthodes statistique et structurelle pour produire des résultats les plus efficaces possibles . Ce genre de traitement est souvent employé dans des projet d'acquisition des informations terminologiques à partir d'un corpus informatisé dans un domaine donnée.

7.4. les outils de reconnaissance automatique des expressions figées

Nous allons présenter dans cette section quelques applications ont pour objectif l'extraction d'information d'ordre terminologique. Ces applications sont connues dans le domaine du TAL et font généralement référence, tels que ACABIT (Daille, 1994) LEXTER (Bourigaulf 1994) et DicAssist

7.4.1. DicAssist

Il s'agit d'un systèmes qui s'appuie sur les ressources de l'Internet pour construire une base de donnée d'expressions figées. Il permet de contrôler toutes les étapes nécessaires à la constitution et la gestion de cette base de données. Ce système à une architecture dite modulaire dans la mesure où elle fait appel à différents serveurs, bases de données et programmes qui créent une unique chaîne de traitement.

En premier lieu, un programme appelé *webget* récupère tous les nouveaux documents édités par un portail donné de la Toile , puis il les enregistre et répertorie dans une base de données . Ensuite, un serveur dénommé SENTA permet de extraire les termes candidats et les enregistrer dans une base de données comportant les expressions figées potentilles qui seront validées manuellement. Ces expressions doivent également être enrichies linguistiquement,en associant chaque expression aux différents types d'expressions figées proposés par Gaston Gross(1996, à savoir c'est-à-dire catégoriser ces expressions en tant que noms composés, locutions verbales, locutions adjectivales, déterminants composés, locutions adverbiales ou locutions prépositives ou conjonctives.

SENTA, acronyme de « Software for the Extraction of N-ary Textual Associations » est un logiciel visant l'extraction des séquence textuelles composée de Nmots (unité lexicales). En effet le fonctionnement de ce logiciel est basé sur trois concepts essentiels:

1- les modèles Ngrams positionnels: Il s'agit de la construction des séquences ordonnées de N unités lexicales correspondant aux énoncés du corpus . Pour ce logiciel le N pourrait aller jusqu'à 7grams

2- "Expectative Mutuelle": un programme qui permet de mesurer si les séquences établies par les modèles de N-grams positionnels construits à partir du texte constituent des expressions figées .Il ne tient pas compte de la fréquence des occurrences des séquences extraites mais c'est traitement statistique permet de mesurer à quel point la présence d'un mot est essentielle pour garantir une interprétation figée dans un N-gram positionnel ou ce qui est appelé *le degré de cohésion* .

3- GenLoclmax: un Algorithme de sélection pour les Ngrams les plus aptes à constituer des séquences figées.

DicAssist est, donc est application constituée d'une chaîne de traitements qui permet de constituer un base de donnée d'expression figées à partir des ressource de la Toile

7.4.2. ACABIT

Il s' agit également de programme d'extraction de terminologie utilisé particulièrement par Béatrice Daille en 2001 dans une recherche concernant les adjectifs relationnels . Ce programme procède à l'extraction des séquences limitées à deux unités lexicales pleines qui correspondent au Nom suivi d'un Adjectif et qui pourrait être plus longue quand le nom est modifié par un groupe prépositionnel. Le corpus ouvert en entrée doit, avant tout traitement être étiqueté et lemmatisé [1], ensuit le programme à recourt à des programme locale à base de d'expressions régulières pour l'extraction des termes candidats .

7.4.3. LEXTER

C'est un logiciel développé par Didier Bourigault en 1994 dans le cadre des recherches d'EDF. D. Bourigault considère que les méthodes d'extraction basées sur des critères de fréquence ne sont pas les plus appropriées et les plus efficaces pour l'extraction de termes complexes, des méthodes à caractère davantage linguistique seraient plus

[1] *L'étiquetage e*t *la lemmatisations* sont deux procédés qui s'effectuent au début du traitement automatique du corpus : L' étiquetage consiste à l'affectation à chaque mots des étiquettes comportant des informations morphosyntaxiques en fonction du contexte au moyen d'un étiqueteur comme *Tree Tagger* . La lemmatisation consiste à remplacer une forme fléchie par son lemme, ce qui permet de définir des formes de base dénuée de toute variation pour chaque mot. Cette opération est assurée par un programme appelé lemmatiseur comme le programme *Flemm* qui lemmatise un fichier étiqueté les étiqueteurs *Bill ou Tree Tagger* et calcule le Lemme à base de centaine de règles et d'u lexique réduit qui comporte près de 3000 mots.

adaptées dans la mesure où elles font appel aux caractéristiques linguistiques et formelles du terme, ce qui permet d'obtenir les résultats les plus précis possibles. Cette méthode se base sur une analyse syntaxique et sur des calculs statistiques appliqués sur la liste des termes candidats pour retenir les termes complexe les plus pertinents .

Nous constatons que ces trois outils sont utilisé, tous pour extraire des terminologies à partir d'un corpus donnée tandis que ces leur fonctionnement se base sur des méthodologies différentes. En effet, l'extraction des termes candidats avec DicAssist se fait indépendamment de toute information linguistique, ces informations sont apportées en aval du traitement. A l'inverse, ACABIT et LEXTER fonde leur fonctionnement sur le traitement linguistique du corpus, à savoir l'étiquetage, la lemmatisation ou encore l'analyse syntaxique du corpus pour procéder ensuite à des traitements statistiques pour affiner les résultats de l'extraction[1]. Par ailleurs, il existe d'autre méthodes pour la reconnaissances automatique des séquences complexes qui s'appuient sur les dictionnaires électroniques regroupant les diverses tables constituées par le lexique-grammaire élaboré par le LADL. Ces dictionnaires et à l'aide des grammaire (sous la forme de graphes) ont été appliqués à des textes avec les logicels *Unitex* (Paumier 2002) et notamment l' application *Intex* et leur extension *Nooj* conçus et développés par Max Silberztein. Cette méthode et plus particulièrement ces deux dernières applications seront étudiées, dans le cadre de cette recherche, et appliquées à un corpus de texte journalistiques dans le chapitre suivant.

[1] LE ROI, Marie-Véronique, Op.cit, p. 45.

IV. Traitement automatique des expressions figées

1. Présentation du corpus.

Nous avons choisi pour notre étude un corpus de textes journalistiques .Il s'agit de différents numéros du journal algérien El watan que nous avons récupéré des archive informatisées du journal. Nous avons opter pour ce types de textes en raison de leur variété discursive et leur richesse lexicales .En effet le texte journalistiques présente plusieurs critères qui favorise sa validité comme un champ d'étude pour le phénomène du figement .

D'abord, la détection de ce phénomène exige un corpus copieux et évolutif, c'est-à-dire, des centaines, voire des milliers de pages d'un discours produit à travers des périodes consécutives .Alors il n'y aurait pas mieux qu'un journal publié quotidiennement. Ensuite le journal se caractérise par la variété , ainsi sur le plan thématique qu' au niveau des registres et des spécialités linguistiques surtout pour les journaux généralistes où on alloue une rubrique à chaque domaines (politique ,économie , société , culture et littérature , science et technologie ,sport et loisirs etc.). Ce genre de texte jouit également d' une *hétérogénéité énonciative* pertinente, dans la mesure où la prise en charge de l'énonciation est d'une altérité explicite*(hétérogénéité montrée*) et implicite (*(hétérogénéité constitutive*)[1].Cette propriété fait que le texte laisse mettre en œuvre les énoncé et les mots de l'autrui, à force d'être mémorisés et réemployer par les locuteurs comme étant une partie du système de la langue ; c'est, ces expressions que l'on considère actuellement comme figée ou en cours de lexicalisation.

2. Prétraitement du corpus

En premier lieu, nous avons téléchargé 16 numéros du journal *El Watan* plus les compléments *économie*, *Tv* et *immobilier* à partir de l'archive électronique du journal. Puis, nous les avons convertis au format "texte" pour que l'on en conserve le contenu textuel seulement et pour que l'on puisse le traiter automatiquement. Voici une page du journal avant la conversion :

[1] CHARAUDEAU, Patrick, MAINGUENEAU, Dominique, *Dictionnaire d'analyse du discours*,Le Seuil, Paris, 2002, p. 292.


El Watan

LE QUOTIDIEN INDÉPENDANT - Vendredi 30 juin - Samedi 1er juillet 2006

BELLOUTA (JIJEL)
Deux gardes communaux assassinés P. 6

32 Pages

ACCOUCHEMENTS ET PRISE EN CHARGE DES NOUVEAU-NÉS
Graves défaillances dans les hôpitaux

LA BIPARTITE UGTA-GOUVERNEMENT SE TIENDRA DEMAIN
LE DOSSIER DES SALAIRES AU MENU

Le début d'un dialogue

Béjaïa surclasse le MC Alger

Haier Service d'Assistance Clients

« LE QUOTIDIEN INDÉPENDANT - Vendredi 30 juin - Samedi 1er juillet 2006 El Watan Deux gardes communaux
assassinés P. 6 3322 Pages
BELLOUTA (JIJEL)
N° 4751 - Seizième année - Prix : Algérie : 10 DA. France : 1 ¤ . USA : 2,15 $. ISSN : 1111-0333 - http://www.elwatan.com
. La 13e bipartite gouvernement-UGTA se tiendra demain
au Palais du gouvernement . Les deux parties discuteront de
l'augmentation des salaires de 1,5 million de fonctionnaires,
oscillant entre 1660 et 5415 DA.
I l a suffi d'une décision du président Bouteflika
pour que les choses se précipitent.
Quelques jours après la déclaration du chef de
l'Etat sur la revalorisation des salaires de la
Fonction publique, le ministère du Travail a
annoncé, hier dans un communiqué de presse,
la tenue, dimanche 2 juillet, de la 13e bipartite
gouvernement - UGTA. Après plusieurs mois
de valse-hésitation autour de la date d'une (…) »[1]

Figure IV.1.Une page du journal Elwatan en format pdf

Et voila la page après la conversion au format « texte » :

Figure IV.2. Un extrait d'une page convertie au format texte

[1] *El Watan* ,du Vendredi 1er juillet 2006 ,p.1

Ensuite, nous chargeons le texte converti dans un logiciel pour en faire le traitement

3. la traitement automatique

Pour appréhender le phénomène du figement nous avons opté pour les méthodes de TAL (*Traitement automatique de la langue),* notamment celles développées par le LADL (voir infra). Cette méthode consiste à étudier les occurrences et les fréquences des suites de mots dans un corpus textuel au moyen des logiciels appropriés à ce genre de traitement. A travers les données obtenues nous pourrions repérer les expressions qui pourraient être figées ou en cours de figement.

3.1. Les outils de traitement

3.1.1. Les dictionnaires électroniques

3.1.1.1. Lexicographie traditionnelle et dictionnaires électroniques

Nous allons en premier lieu, étudier ce qui distingue la lexicographie traditionnelle des dictionnaires électroniques Dans la tradition,on divise généralement le dictionnaire et la grammaire qui représentent les outils normatifs indispensables pour décrire la langue et le bon usage. Le dictionnaire qui a pour objectif la description du lexique permet de recenser toutes les irrégularités d'une langue tandis que la grammaire établirait des règles décrivant la régularité , néanmoins nous constations les dictionnaire font appel à ces deux descriptions qui sont finalement davantage liées que totalement opposées.

L'élaboration de dictionnaires repose cependant fondamentalement sur la distinction entre le lexique et la syntaxe. D'après Gaston Gross, l'analyse syntaxique devrait précéder toute démarche lexicographique. Le but premier d'un dictionnaire est la description des mots ou unités lexicales qui composent une langue. L'objectif serait plus précisément de donner le « sens » d'un mot. Le sens d'un énoncé constitué d'un certain nombre de mots résulterait donc de la somme des mots composant cet énoncé. Les dictionnaires ont généralement recours à la syntaxe pour illustrer les acceptions des mots décrits. Une notion importante dans les dictionnaires concerne les catégories associées aux mots du dictionnaire. Gaston Gross dit par ailleurs que « tout dictionnaire repose sur la notion de catégorie »[1].

La lexicographie n'exclut donc pas totalement les informations d'ordre syntaxique. Cette question de la séparation lexique/syntaxe se pose particulièrement pour la description des séquences figées. Dans le cas des locutions verbales, la question peut effectivement se

[1] LEROI, Marie Véronique, Op.cit, p 47.

poser : une locution verbale doit-elle figurer dans un dictionnaire ou dans un ouvrage de grammaire ? En effet, ces séquences se trouvent à la limite de ces deux disciplines dans la mesure où elles sont régies par des règles syntaxiques régulièrement mais qu'elles sont sémantiquement équivalentes à une unique unité lexicale. Si une locution verbale est consignée dans un dictionnaire, sous quelle entrée doit-elle être décrite ? Est-ce sous l'entrée correspondant au verbe ou celle correspondant au nom qui constitue la tête du groupe nominal objet ? G. Gross (*1987*) estime que la solution adoptée par les dictionnaires n'est pas satisfaisante dans la mesure où les expressions figées figurent généralement en fin d'article pour en souligner un emploi figuré. G. Gross répond cependant à une de ces questions en indiquant que les expressions figées n'étant pas prédictibles doivent être décrites dans le dictionnaire : le sens de ces séquences ne pouvant être obtenu par celui des éléments constituants, ces expressions doivent faire l'objet d'une mémorisation comme lorsque l'on apprend un nouveau mot.

Un dictionnaire d'usage contient donc principalement des informations correspondant au lemme du mot, à sa catégorie syntaxique, à sa définition, à ses différentes acceptions et éventuellement quelques exemples.

Les dictionnaires électroniques sont généralement constitués d'une manière toute autre. En effet, les dictionnaires électroniques ne sont normalement pas destinés à être utilisés par un utilisateur humain qui se substitue à un ordinateur-utilisateur. Ces dictionnaires, dans cette perspective, ne constituent donc que des ressources d'une utilité précieuse qui permettent l'exécution d'un programme informatique.

Ces dictionnaires se présentent sous la forme de bases de données lexicales entièrement formalisées afin d'éviter toute ambiguïté lors d'un traitement automatique. Les propriétés lexicales définissant chaque entrée comportent les informations les plus précises et explicites possibles afin d'éviter l'échec de la reconnaissance automatique.

Les dictionnaires électroniques différent donc en de nombreux points des dictionnaires classiques. En effet n'étant pas destinés à être utilisés par un être humain, les dictionnaires électroniques doivent donc être aussi complets que possible. et les informations données par ces dictionnaires doivent également être explicites :tandis que les dictionnaires classiques n'ont généralement pas la nécessité d'être explicites car ils font appel aux connaissances pragmatiques et à l'adaptabilité des utilisateurs.

Les informations fournies pas les dictionnaires électroniques sont totalement codées et exsangues d'information d'ordre sémantique, à savoir d'indication de sens. Ces informations sont en effet destinées à être exploitées par des programmes informatiques

afin de procéder à des traitements dans les divers secteurs du TAL.

Les caractéristiques que nous venons de définir s'appliquent particulièrement pour les travaux du LADL., les résultats émanant des travaux du LADL sont représentés sous forme de tables qui représentent le lexique-grammaire, par ailleurs des dictionnaires électroniques ont aussi été élaborés par le TAL. Ces dictionnaires électroniques constituent des ressources sur lesquelles s'appuient des outils d'analyse ou d'acquisition de termes. Les divers dictionnaires du LADL renvoient plus exactement aux dictionnaires DELA (Dictionnaire Electronique du LADL). La construction de ces dictionnaires est basée sur une définition purement formelle du mot simple, qui diffère totalement de la description morphologique.

3.1.1.2. Les unités lexicales

Les unités lexicales sont les unités linguistiques qui constituent les atomes de la phrase, c'est-à-dire les unités non-analysables de la langue. Intuitivement, ce sont des mots dont le sens ou la fonction grammaticale ne peuvent pas être calculés : il faut les apprendre pour pouvoir les utiliser. D'un point de vue linguistique, il est indispensable de les recenser en extension et de décrire leurs propriétés (généralement dans un lexique, d'où le nom d'*unités lexicales*).

D'un point de vue formel et d'après Max Silberztein, on peut classer les unités lexicales en quatre classes :

> *« les mots simples **(Simple Words)** sont les unités linguistiques atomiques qui s'écrivent sous la forme de formes simples. Par exemple, pomme, table.*
> *Les morphèmes **(Morphemes)** sont les unités linguistiques atomiques qui sont des séquences de lettres inclues dans des formes simples. Par ex. : re-, -ation.*
> *Les mots composés **(Compound Words)** sont les unités linguistiques atomiques qui sont des séquences de lettres et de séparateurs. Par exemple. : cordon bleu, pomme de terre.*
> *Les expressions figées **(Frozen Expressions)** sont les unités linguistiques atomiques qui sont des, séquences potentiellement discontinues de lettres et de séparateurs. Par ex. : prendre … en compte, ne … pas. »*[1]

[1] SELBERZTEIN, Max, *Intex* , www.mshe.univ-fcomte.fr/intex/downloads/Manuel.pdf, consulté le 11-03-20008.

Un mot simple se réduit donc à une unité de texte définie sur l'alphabet des codes ASCII et ne comportant aucun séparateur (ni trait d'union, ni blanc ni apostrophe). Cette définition tient donc fortement compte de la graphie des unités. L'alphabet des codes ASCII compte plus de vingt-six lettres dans la mesure où il comprend également les divers caractères accentués disponibles en français. Cette définition du mot simple aboutit à la définition suivante du mot composé : « un mot composé est une séquence de mots simples». Les mots composés sont à distinguer des groupes libres de mots simples. Silberztein prend pour illustrer ce point de vue les deux exemples suivants :

Cordon rouge et cordon bleu.

- *cordon rouge :* (cordon de couleur rouge) est un groupe libre de mots simples.
- *cordon bleu* (bon cuisinier) est un mot composé

Le principe permettant de dissocier ces deux types de mots est le suivant :

Une séquence de mots simples est figée (ou composée) si l'une au moins de ses propriétés syntaxiques, distributionnelles ou sémantiques ne peut être déduite des propriétés de ses constituants.

3.1.1.3. Les dictionnaires du LADL

Les différents dictionnaires électroniques du LADL appelés conventionnellement DELA (Dictionnaire Electronique Du LADL) sont au nombre de quatre. Le DELAS décrit donc la morphologie et la flexion des mots simples ; le DELAC a pour objet la description des mots composés ; le DELAF décrit les formes fléchies et les lexiques dérivés du français ; le DELACF est généré automatiquement à partir du DELAC pour décrire les formes fléchies et composées du lexique ces différents dictionnaires électroniques constituent les ressources sur lesquelles s'appuient des logiciels tels que Intex dont nous décrirons le fonctionnement dans une prochaine section. Par ailleurs ces dictionnaires que nous venons de décrire ne sont cependant pas les seuls disponibles. En effet, les versions numérisées des dictionnaires classiques initialement sur support papier reçoivent également l'appellation de dictionnaires électroniques. La version informatisée du TLF, autrement dit *le Trésor de la langue Française,* par exemple, est un dictionnaire électronique qui peut également constituer une ressource pour des programmes en TAL.

3.1.1.3.1. les dictionnaires DELAF

Les programmes informatique comme INTEX que nous allons utiliser dans notre recherche, ont besoin, pour traiter des textes écrits dans une langue donnée , d'un dictionnaire qui reconnaît, recense et décrit les mot simple de la langue, avec leurs formes

fléchés (en français : les formes conjuguées pour les verbes, formes au pluriel et au féminin pou les noms et les adjectifs)

Les dictionnaires DELAF peuvent être construits automatiquement à partir de dictionnaires de types DELAS qui ne contiennent que les lemmes des formes c'est-à-dire (l'infinitif des verbes, le masculin singulier pour les nomes et les noms et les adjectifs)

Voici par exemple quelques entrées du DELAF des langues françaises:

« *avions,avion.N+Conc:mp*
avions,avoir.V+aux:I1p:S1p
cousins,cousin.N+Anl:mp
cousins,cousin.N+Hum:mp
de,de.PREP

*La première ligne représente le fait que la forme avions est associée au lemme avion ;c'est un nom (**N**), dont la classe distributionnelle est Concret (**+Conc**) ; la forme représente le masculin pluriel (**:mp**). La seconde ligne représente le fait que la forme avions est aussi associée au lemme avoir, qui est un verbe (**V**) auxiliaire (**+aux**) ; la forme est conjuguée à l'imparfait, première personne du pluriel (**:I1p**) ou au subjonctif présent, première personne du pluriel (**:S1p**). Les deux lignes suivantes représentent les deux formes nominales cousins (animal ou humain). La dernière ligne représente la forme de, qui est identique à son lemme, et est une préposition (**PREP**). »*[1]

Dans le logiciel INTEX nous trouvons deux versions des dictionnaires au format DELAF avec codes morphologiques seulement conçus par Blandine Courtois,2002 .:

-*Delaf.bin :*746 214 entrées ; 35 822 différentes.

- *Delafm.bin* : 682 457 entrées ; 7 785 différentes.

Cette application dispose, aussi de plusieurs autres dictionnaires spécialisés conçus par Max Silberztein à savoir :

-un dictionnaire de prénoms (environ 500 entrées) ; en voici deux entrées :

Abraham,.N+PR+Hum:ms

Dominique,.N+PR+Hum:ms:fs

-un dictionnaire des noms propres (*Noms propres +.fst)*

-un dictionnaire des organisations et de compagnies (*Organisations .dic)*

[1] Ibid, p.106.

-un dictionnaire des sigles ex ADN, CSA (Sigles.dic)

-un dictionnaire des noms des pays et des régions (Toponymes.dic)

-un dictionnaire contenant quelques abréviations ex :Mme=Madame (*Abréviation dic.)*

-un dictionnaire contenant quelques noms propres de personnalités *(célebrités. dic)*

-un dictionnaire contenant quelques noms propres de personnalités politiques *(célebrités politiques.dic)*

-un petit dictionnaire qui sert à cacher des entrées du DELFA.

Quant à l'application NOOJ que nous allons , aussi employer au cous de cette recherche ,elle contient, en outre des ces dictionnaires :

-un dictionnaire qui prend en charge les formes élidés (*Elisions,nod*).

-un dictionnaire pour décrire des mots invariables ,des variations orthographiques , des noms et des verbes associés à des paradigmes flexionnels et dératisations et les noms composés qui se fléchissent.*(Exemples.nod*)

- un dictionnaires spécialisé pour les noms des métiers.(métiers .nod)

3.1.1.3.2. les dictionnaires électroniques de types DELACF

Ce sont ces dictionnaires qui nous intéressent le plus puisque il s'agit des actionnaires des mots composés dont les entrées contiennent plus de mots simples Il. défère des dictionnaires DELAF en ce que les entrées lexicales(le texte en début de la ligne et avant la virgule) et ou les lemmes (le texte entre la virgule et le point) peuvent contenir des séparateurs .et comme tous les dictionnaires électroniques destinés à être employer par un programme informatique, ils ne fournissent que des descriptions morphologiques et syntaxiques de ces séquences qui servent à leur reconnaissance automatique et d'autres éventuels traitement informatique . voici par exemple quelques entrées du dictionnaire DELACF du français :

cousins germains,cousin germain.N+NA+Hum:mp

criant de vérité,.A+EPN:ms

pommes de terre,pomme de terre.N+NDN+Conc:fp

tant et si bien que,.CONJS+3

tout à coup,tout à coup.ADV+PCPC[1]

[1] Ibid, p. 111.

3.1.1.3.2.1. Difficulté de reconnaissance : Groupes nominaux libres vs mots composés

Dans l'analyse automatique de textes, nous confrontons le problème de la limite entre noms composés (qu'il faut lexicaliser) et groupes nominaux libres. En effet il existe plusieurs centaines de milliers de cas plus difficiles, comme, par exemple :carte routière, ceinture noire, machine à laver qui font que certaines applications fourniront des résultats incorrects.

De ce fait et dans le cadre des travaux du LADL , Max Silberztein définit trois critère principales pour dessiner la entre les groupes nominaux que l'on doit lexicaliser, et ceux qu'on ne veut pas lexicaliser :

a) Atomicité sémantique : si le sens exact du groupe nominal ne peut pas être déduit du sens de ses composants, le groupe nominal doit être lexicalisé et il est donc traité comme un nom composé. Max Silberztein pense que le groupe nominal *carte routière* doit être lexicalisé piusque le mot carte n' pris dans son sens générale (*carte depaiment, carte d'abonnement, carte d'identification, carte électronique, carte de visite*, etc.) mais, seulement dans le sens de *carte* (géographique (qui sert à (se) localiser) ; contrairement aux groupes nominaux *carte ancienne* ou *carte plastifiée* qui sont ambigus

b) Restriction distributionnelle : certains des constituants du groupe nominal qui appartiennent par ailleurs à des classes distributionnelles naturelles, ne peuvent pas être remplacés de façon libre,

Par exemple, seuls sept adjectifs de couleur peuvent se combiner avec *ceinture* dans un groupe nominal qui peut représenter un nom humain : « *Luc est ceinture noire* ». Ces adjectifs de couleur ne sont pas prévisibles a priori ; ils ne sont pas identiques aux adjectifs que l'on trouve dans « *Luc est maillot jaune* » ou « *Luc est un col blanc* ». En conséquence, il faut recenser ces sept groupes nominaux, ce qui revient à les traiter comme des noms composés.

c) Institutionnalisation de l'usage : l'utilisation de certains groupes nominaux, est et quasi-obligatoire, et on peut pas les remplacer par d'autres constructions syntaxiques potentielles tout aussi valides, Silberztein estime que ces groupes nominaux représentent la façon « institutionnalisée » de représenter des objets ou des concepts, ce qui revient à les traiter comme des noms composés.

Par exemple, la langue française nous permettrait a priori de nommer les *machines àlaver* d'au moins une douzaine de façons :

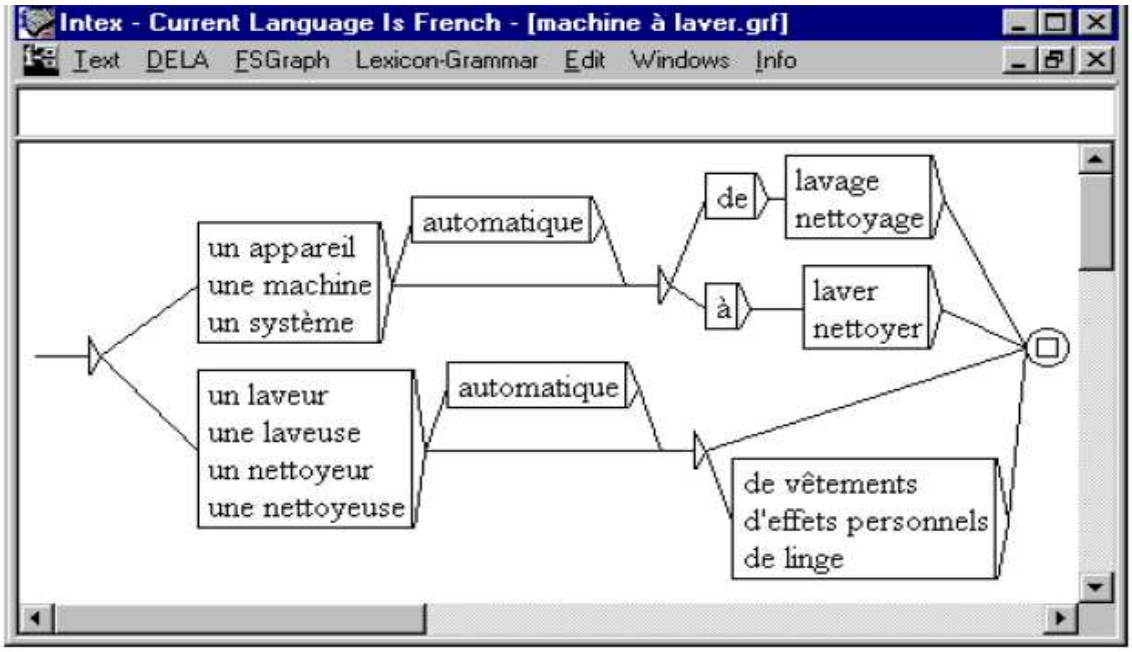


*Figure IV.3.*Un graphe représentant quelques variantes syntaxiques correctes du nom composé *machine à laver*[1]

Même si la langue française autorise tous ces groupe nominaux , seuls trois sont utilisés : *un lave-linge, une machine à laver, une machine à laver le linge*. De ce fait, Il est essentiel de recenser ces trois groupes nominaux particuliers, pour éviter des traductions maladroites (*washingmachine => machine lavante*), pour indexer correctement les documents (par exemple *machine à laver* doit être indexé avec *lave-linge*, mais pas avec *la vaisselle*), pour des applications pédagogiques (enseignement du français seconde langue) et plus généralement pour décrire le vocabulaire de la langue .Ces trois critères définissent un ensemble de mots composés bien plus grand que celui généralement admis. Les estimations qui ont pu être faites au LADL montrent que pour couvrir le vocabulaire standard du français, environ 250 000 mots composés doivent être recensés et décrits.

3.1.1.3.2.2. Utilisations particulières des DELACF

1-le « lemme » qui est la forme canonique de l'entrée donne à la forme composé une variante plus langue et explicite : associer à chaque forme composée une variante plus longue et explicite ; par

Exemple :

Etats-Unis,Etats-Unis d'Amérique. N+Géo+Pays:mp

[1] Ibid, p. 113.

U.S.A.,Etats-Unis d'Amérique.N+Géo+Pays:mp

USA,Etats-Unis d'Amérique.N+Géo+Pays:mp

Inversement, le concepteur de dictionnaire de termes peut choisir comme forme canonique un mot simple *:*

roman policier,polar. N+Conc:ms

roman policier de la série noire,polar. N+Conc:ms

Les dictionnaires de mots composés peuvent aussi contenir des entrées qui sont en fait des mots simples, par exemple :

carte,carte bancaire. N+Conc:fs/carte bleue

carte,carte géographique. N+Conc:fs/carte routière

carte,carte électronique. N+Conc:fs/carte-mère

carte,carte postale. N+Conc:fs

Cette utilisation des DELACF permet de lever l'ambiguïté que présentent certains mots simples, comme le cas du mot carte , ce qui sert ,aussi à donner une traduction correcte :

carte bancaire, credit card. N+Conc:fs

carte bancaire, debit card. N+Conc:fs

Pour l'application INTEX le dictionnaire DELACF français est actuellement disponible en quatre fichiers :

-Le fichier **Noms.bin** contient 244 726 formes de noms composés, répartis en une dizaine de classes qui correspondent à la structure morpho-syntaxique des entrées : par exemple, **+NA** pour *Nom Adjectif,* **+NDN** pour *Nom de Nom,* etc. Cf. B. Courtois, M. Silberztein Eds 1990 pour une description des types de noms composés ;

-Le fichier **Adv.bin** contient 7 753 adverbes figés, soit dans une forme qui a fonction d'adverbe (ex. *dans l'intimité la plus stricte*), soit dans une forme qui a fonction de préposition (ex. *à grand renfort de*) ; chaque entrée est associée à la table du lexiquegrammaire dans laquelle les propriétés syntaxiques de l'adverbe correspondant sont décrites (cf. M. Gross 1986) ;

-Le fichier **EPN.bin** contient 14 551 formes figées utilisées avec le verbe *être,* dans des formes à fonction adjectivale (ex. *criant de vérité*), adverbiale (ex. *à un epsilon près*) ou de préposition (ex. *une planche de salut pour*). Cf. M. Gross 1997 pour une description des **EPN** ;

-Le fichier **Conjs.bin** contient 1 591 conjonctions de subordination, soit dans une forme qui a fonction de conjonction (ex. *tant et si bien que*), d'adverbe (ex. *dans ce cas*) ou de préposition (ex. *à défaut de*).
INTEX contient aussi pour le français quelques exemples de dictionnaires de mots composés, comme par exemple un dictionnaire de pronoms (ex. *quelques-uns*), de toponymes (ex. *New-York, Afrique du Sud*), de prénoms (ex. *Jean-Paul*), etc.[1]

3.1.1.3.3. Le Dictionnaire Explicatif et Combinatoire (DEC)

Par ailleurs Igor Mel'cuk avec la collaboration d'Alain Polguère et André Clas. dans une démarche lexicologique ont conçu un dictionnaire dit Explicatif et combinatoire (DEC) qui défère des dictionnaire classique issus de la lexicographie. Il s' agit d'une théorie dite *sens-texte* qui tente de produire à un dictionnaire contenant toutes les informations qui pourraient permettre à un locuteur de construire toutes les expressions linguistiques correctes de n'importe quelle pensée et ce, dans n'importe quel contexte. En effet cette approche proposer de partir des représentations sémantiques de la langue, autrement dit les sens et les acceptions à l'aide du lexique. Le lexie , pour cette théorie correspond à chaque acceptions possible d'un mot simple ou d'une locution .Si un mot donné est polysémique, le nombre de lexies disponibles pour ce même mot correspondra au nombre d'acceptions que ce mot reçoit. Donc l'acceptions du terme *lexie* ne correspond pas à celle de Bernard Pottier (voir infra).

Chaque lexie est décrite dans le DEC selon sa définition, ses connotations, et d'autres informations qui n'apparaissent pas ou alors apparaissent succinctement dans un dictionnaire classique. De ce fait Un article du dictionnaire qui décrit la lexie doit comprendre dix zones principales et un tableau pour les illustrer :

[1] LEROI, Marie Véronique,Op.cit, p. 50

ARTICLE : /LEXIE/		
1	Zone vedette	- Lexie vedette - Variante orthographique
2	Zone phonologique	- Prononciation - Prosodie particulière
3	Zone morphologique	- Partie du discours (ou catégorie) - Type de déclinaison ou de conjugaison - Formes irrégulières ou non réalisables
4	Zone stylistique	- Marques d'usage
5	Zone sémantique	- Définition - Connotations
6	Zone de combinatoire syntaxique	- Restrictions sur la cooccurrence syntaxique
7	Zone de combinatoire lexicale restreinte	- Restriction sur la cooccurrence lexicale
8	Zone d'exemples	- Exemples
9	Zone phraséologique	- Emplois figés
10	Zone de Nota Bene	- Remarques diverses

Tableau.IV.1.Les différents zones de description constituant un article du DEC[1]

De ce fait des locutions telle que « se mettre le doigt dans l'oeil » constitue une entrée de dictionnaire explicatif et combinatoire ,au même titre que « se fourrer le doigt dans l'oeil ».

La version électronique de ce dictionnaire est en cours de réalisation en informatisant e les quatre volumes du DEC

Cette manière de conception de dictionnaire permet, donc d'intégrer d'une manière plutôt satisfaisante les locutions verbales et d'autres séquences figées et d'envisager une nouvelle lexicographie phraséologique .

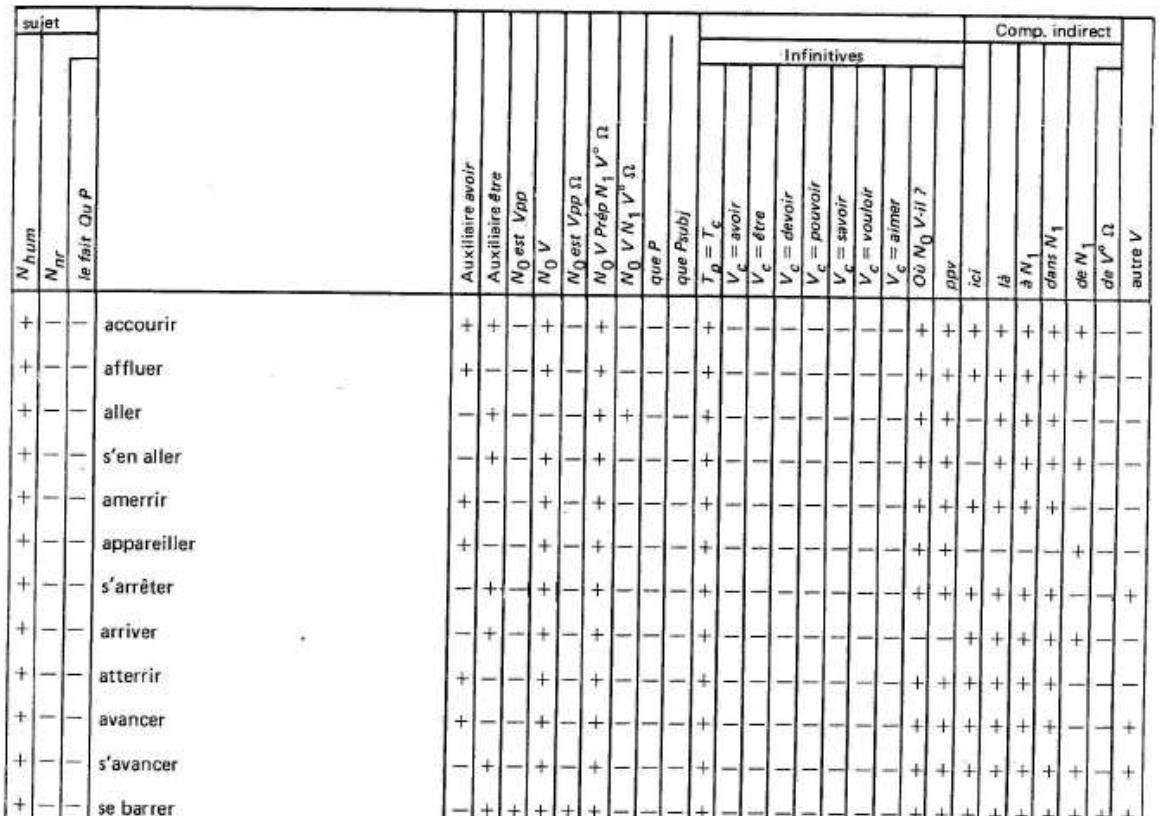
3.1.2. les tables de lexique-gammaire

3.1.2.1. la constitution de la table de lexique-grammaire

Les tables de lexique-grammaire sont un moyen simple et efficace de représenter le comportement distributionnel et transformationnel des prédicats dans les phrases simples Elles ont la forme de matrices. Chaque ligne correspond à une entrée lexicale (ou un

[1] LEROI, Marie Véronique,Op.cit, p.51.

prédicat). Chaque colonne correspond à une propriété. A l'intersection, il y a un signe + si l'entrée lexicale accepte cette propriété ; un signe – si elle ne l'accepte pas. Voici un exemple d' une table de lexique-grammaire :



sujet				Auxiliaire avoir	Auxiliaire être	N_0 est Vpp	N_0 V	N_0 est Vpp Ω	N_0 V Prép N_1 V° Ω	N_0 V N_1 V° Ω	que P	que P_{subj}	Infinitives									Comp. indirect						
													$T_0 = T_c$	V_c = avoir	V_c = être	V_c = devoir	V_c = pouvoir	V_c = savoir	V_c = vouloir	V_c = aimer	Où N_0 V-il ?	ppv	ici	là	à N_1	dans N_1	de N_1	de V° Ω	autre V
N_{hum}	N_{nr}	le fait Qu P																										
+	–	–	accourir	+	+	–	+	–	+	–	–	–	+	–	–	–	–	–	–	–	+	+	+	+	+	+	–	–
+	–	–	affluer	+	–	–	+	–	+	–	–	–	+	–	–	–	–	–	–	–	+	+	+	+	+	+	–	–
+	–	–	aller	–	+	–	–	–	+	+	–	–	+	–	–	–	–	–	–	–	+	+	–	+	+	–	–	–
+	–	–	s'en aller	–	+	–	+	–	+	–	–	–	+	–	–	–	–	–	–	–	+	+	–	+	+	+	–	–
+	–	–	amerrir	+	–	–	+	–	+	–	–	–	+	–	–	–	–	–	–	–	+	+	+	+	+	–	–	–
+	–	–	appareiller	+	–	–	+	–	+	–	–	–	+	–	–	–	–	–	–	–	+	+	–	–	–	+	–	–
+	–	–	s'arrêter	–	+	–	+	–	+	–	–	–	+	–	–	–	–	–	–	–	+	+	+	+	+	–	–	+
+	–	–	arriver	–	+	–	+	–	+	–	–	–	+	–	–	–	–	–	–	–	–	–	+	+	+	+	–	–
+	–	–	atterrir	+	–	–	+	–	+	–	–	–	+	–	–	–	–	–	–	–	+	+	+	+	+	–	–	–
+	–	–	avancer	+	–	–	+	–	+	–	–	–	+	–	–	–	–	–	–	–	+	+	+	+	+	–	–	+
+	–	–	s'avancer	–	+	–	+	–	+	–	–	–	+	–	–	–	–	–	–	–	+	+	+	+	+	+	–	+
+	–	–	se barrer	–	+	+	+	+	+	–	–	–	+	–	–	–	–	–	–	–	+	+	+	+	+	+	–	+

Tableau.VI.2.Une table de lexique-grammaire[1]

Ces tables regroupent les usages des mots prédicatifs qui partagent les propriétés dites définitoires de la table. En particulier, toutes les entrées d'une table ont en commun un (parfois deux) cadre(s) de souscatégorisation de base. Pour chaque lemme d'une table, les colonnes indiquent en outre des propriétés de sous-catégorisation additionnelles pour ce lemme et en particulier des informations sur :

– les réalisations possibles des arguments (catégorie, préposition, complémenteur, etc.) ;

– les propriétés syntaxiques du verbe ou de ses arguments (réflexivisation, clitisisation,etc.)

– les sous-catégorisations alternatives ;

– les possibilités de redistributions (passif long, passif court, etc.).

Voila deux lignes de table 8 :

[1] CLAIRE,Gardent, BRUNO, Guillaume, et all, *lexique syntaxique et tables du LADL*, CNRC/LORIA, Nancy, France 2006, www.exsynt.inria.fr/talk/parisSept05.pdf , consulté le 16-06-2008.p.237.

N0 =; Nhum	N0 =; Nnr	N0 =; le fait Qu P	N0 = V1 W	[extrap]		8	N0 est V-ant	N0 V	N0 est Vpp W	N1 =; Qu P	N1 =; Qu Psubj	[pc z.]	N1 =; si P ou si P	N1 = V0 W	Tc =; futur	Tc =; passé	Vc =; devoir	Vc =; pouvoir	Vc =; savoir	N1 =; ce(ci+là)	N1 =; Ppv	de N1 = de là	N1 =; Nhum	N1 =; N-hum	N1 =; le fait Qu P	Prép Nhum = Ppv	[extrap][passif]	de N1 V N0	N0 V contre Nhum
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
+	-	-	-	-	s'	abstenir	+	+	-	-	+	-	-	+	-	-	-	-	-	+	+	-	-	+	-	+	-	-	-
+	-	-	-	-		abuser	-	+	-	+	+	-	-	+	+	-	-	+	+	+	+	-	+	+	+	+	+	-	-

Tableau.IV.3.*Deux ligne de la table 8*[1]

3.1.2.2. Structure d'une table de lexique-grammaire

Les colonnes des tables du LADL sont reliées entre elles par des relations de conjonction, de disjonction et de dépendance. Ainsi, pour la table 8 par exemple :

– les colonnes 13 et 14 *dépendent* des colonnes 11 et 12 (la possibilité d'avoir comme complément une complétive sans préposition ou une proposition interrogative dépend de la possibilité d'avoir un complément phrastique avec le mode indicatif ou subjonctif).

– les colonnes 16 et 17 donnent une information *disjonctive* sur les valeurs de traits atomiques (le complément à l'infinitif est compatible avec un adverbe indiquant le futur, avec un adverbe indiquant le passé, avec les deux ou avec aucun des deux).

– la colonne 2 fournit une information *disjonctive* sur la réalisation de l'argument. (le sujet est "non restreint", c'est-à-dire qu'il peut être un sujet humain, une infinitive ou une complétive).

– les colonnes 6 et 7 fournissent une information *conjonctive* sur le verbe (d'une part le lemme et d'autre part la possibilité d'avoir une particule réflexive "*se*" ou "*s'*").

En fait , cette ressource contient des informations de sous-catégorisation qui sont à la fois détaillées et extensives. Ainsi chaque usage d'un lemme est associé à une description de l'ensemble de ses cadres de sous-catégorisation, et pour un cadre donné, chaque argument est associé à une description de ses propriétés morphosyntaxiques et grammaticales. De plus, le lexique-grammaire couvre non seulement les verbes (environ 6 000 lemmes) mais également les adjectifs et les noms (environ 25 000 constructions à verbe support avec tête nominale ou adjectivale).

[1]CLAIRE,Gardent, BRUNO, Guillaume et all, *Extraction d'information de sous-catégorisation à partir des tables du LADL*, www.loria.fr/~perrier/taln06.pdf , consuté le 16-06-2008, Nancy 2006, p. 3.

Pour que l'on puisse employer ces tables dans les traitement automatiques, ils doivent être gérer par une graphe d'extraction et de reconnaissance de lexique . Le LADL a déjà élaboré graphes pour les tables 1,2,4 et 5, ce que fait 10002 entrées . En outre des graphes pour les tables 7,8,10,11,13,14,16 sont en cours de réalisations . [1] Ces derniers représentent à leur tour une troisième ressource pour les traitement automatiques ,surtout la table Cd1 et son graphe sur lequel nous allons s'appuyer pour l'extraction des expressions figées de notre corpus .

3.1.3. les grammaire (les graphes)

Enfin, les grammaires ou les transducteurs présentées sous forme de graphes sont au départ des extensions des dictionnaires de mots composés puis ont évolué vers des niveaux d'analyse supérieurs (voir la section prochaine). Elles peuvent, par exemple, être éditées à l'aide des éditeurs de graphe de l'application Intex et d'Unitex ue nous allons bien étudier dans la section prochaine . Elles présentent de nombreux avantages indéniables comme la représentation compacte de descriptions linguistiques fines [2].

3.2. L' application informatique "Intex"

Comme nous avons vu dans le chapitre précédent , de nombreux applications informatiques ont été développées dans le domaine du TAL et plus particulièrement, dans les études sur les expressions figées . Nous avons opté dans cette recherche pour la méthode qui s'appuie sur les ressources lexicales électroniques, notamment le logiciel Intex .

3.2.1. Description

Intex est un logiciel créé par le LADL afin de procéder à l'analyse de corpus d'un volume important et langues différentes. Ce logiciel a été développé par Max Silberztein en 1993. Intex est donc un environnement de développement linguistique permettant d'analyser morphologiquement et syntaxiquement un texte afin de procéder à divers traitements. Ces .traitements peuvent consister à rechercher des séquences de diverses natures telles que des lettres, des lexèmes, ou des catégories morphologiques.

Ce logiciel fournit également des outils pour décrire la morphologie flexionnelle et

[1] Ibid , p. 19.
[2] GROSS Maurice ., *The construction of Local Grammars*, In: E. ROCHE and Y. SCHABES (Eds.), Finite State Language Processing, The MIT Press, Cambridge, Mass. pp.329–352, 1997

dérivationnelle, la variation orthographique et terminologique. Le vocabulaire est également décrit qu'il s'agisse de mots simples, de mots composés ou d'expressions figées. Des phénomènes dits « semi-figés » sont également recensés dans le logiciel. L'indexation des mots ou d'expressions figées est possible dans le cadre de l'application. Intex permet également l'accès à des concordanciers ou à des outils permettant l'étude statistique des résultats produits.

3.2.1.1. La notion du transducteur

Les textes ou corpus ouverts en entrée, les dictionnaires électroniques sur lesquels sont basés les analyses ou les grammaires locales utilisées sont représentées à un moment donné du traitement par des *transducteurs* à états finis. Les transducteurs à états finis sont des graphes qui représentent un ensemble de séquences en entrée et leur associe des séquences produites en sortie. Un transducteur peut être un automate à état fini. Un automate à état fini, dit aussi automate fini, est un type particulier de transducteur à état fini. La principale différence qui distingue ces deux procédés consiste dans le fait que le transducteur comporte aussi bien une bande de lecture qu'une bande d'écriture tandis que l'automate à état fini comporte uniquement une bande de lecture et ne permet donc pas de production.

Intex a recours à des expressions régulières pour procéder à la recherche de *motifs* dans les corpus ouverts en entrée. Les graphes qui représentent visuellement les transducteurs à état fini permettent de présenter de manière plus compacte des expressions régulières visant la recherche de motifs complexes. Le transducteur d'une grammaire représente des séquences de mots du texte et fournit des informations linguistiques d'ordre syntaxique sur ces séquences. Le transducteur d'un dictionnaire représente des séquences de lettres qui correspondent aux entrées des unités lexicales et fournit des informations lexicales sur ces séquences. Le transducteur du texte représente des séquences correspondant aux mots qui constituent une phrase du texte. Ces trois objets, dictionnaires, texte et grammaires, ont donc recours au même mode de représentation, ce qui facilite donc l'implémentation du programme.

> *« Une caractéristique essentielle d'INTEX est que tous les objets traités textes,dictionnaires, grammaires) sont à un moment ou à un autre représentés par des* ***transducteurs à états finis****.*
>
> *Un transducteur à état fini est un graphe qui représente un ensemble de séquences en entrée, et leur associe des séquences produites en sortie.*

Typiquement, une grammaire présentera des séquences de mots (lues dans le texte), et produira des informations linguistiques (par exemple des informations sur la structure syntaxique) ; un dictionnaire représentera des séquences de lettres (qui épèlent chaque entrée lexicale), et produira des informations lexicales (partie du discours, codes flexionnels, etc.) ; le transducteur d'un texte présentera les séquences de mots (qui forment chaque phrase) et leur associe des informations lexicales et/ou syntaxiques (les marques linguistiques produites par les différentes analyses). »[1]

La notion de transducteur et d'automate est donc essentielle pour comprendre le fonctionnement d'Intex.

3.2.1.2. Les ressources linguistiques

. Le fonctionnement d'Intex s'appuie sur l'exploitation de trois types de ressources linguistiques :

Parmi ces ressources nous retrouvons donc *les dictionnaires DELA* élaborés par le LADL que nous avons décrit dans la section consacrée aux dictionnaires électroniques. Ces dictionnaires, recensent aussi bien les mots simples que les mots composés.

Les graphes produits par Intex correspondant donc aux dictionnaires, aux grammaires ou aux textes constituent également une de ces ressources linguistiques et présentent l'avantage de présenter de manière compacte des phénomènes linguistiques tant au niveau orthographique, morphologique, que syntagmatique ou syntaxique transformationnel.

Le troisième type de ressource linguistique sur laquelle repose le fonctionnement d'Intex est constitué par *les tables du lexique-grammaire* qui sont comme nous l'avons vu des bases de données qui fournissent une description détaillée des phénomènes linguistiques qui sont à la frontière des disciplines de la syntaxe et du lexique.

3.2. 2. Fonctionnement

3.2.2.1. les information statistiques du texte

Une fois que l'on a démarré Intex, la première étape de traitement d'un corpus consiste à charger un texte en passant par le menu Text > Open … Avant de pouvoir être chargé dans l'application, le texte doit être prétraité et transformé selon des normes propres à Intex et définies par des transducteurs spécifiques. Le chargement du texte s'accompagne par l'apparition d'informations statistiques et formelles concernant ce texte.

[1] SILBERZTREIN, Max, *INTEX*. p. 9. www. mshe.univ-fcomte.fr/intex/downloads/Manuel.pdf.

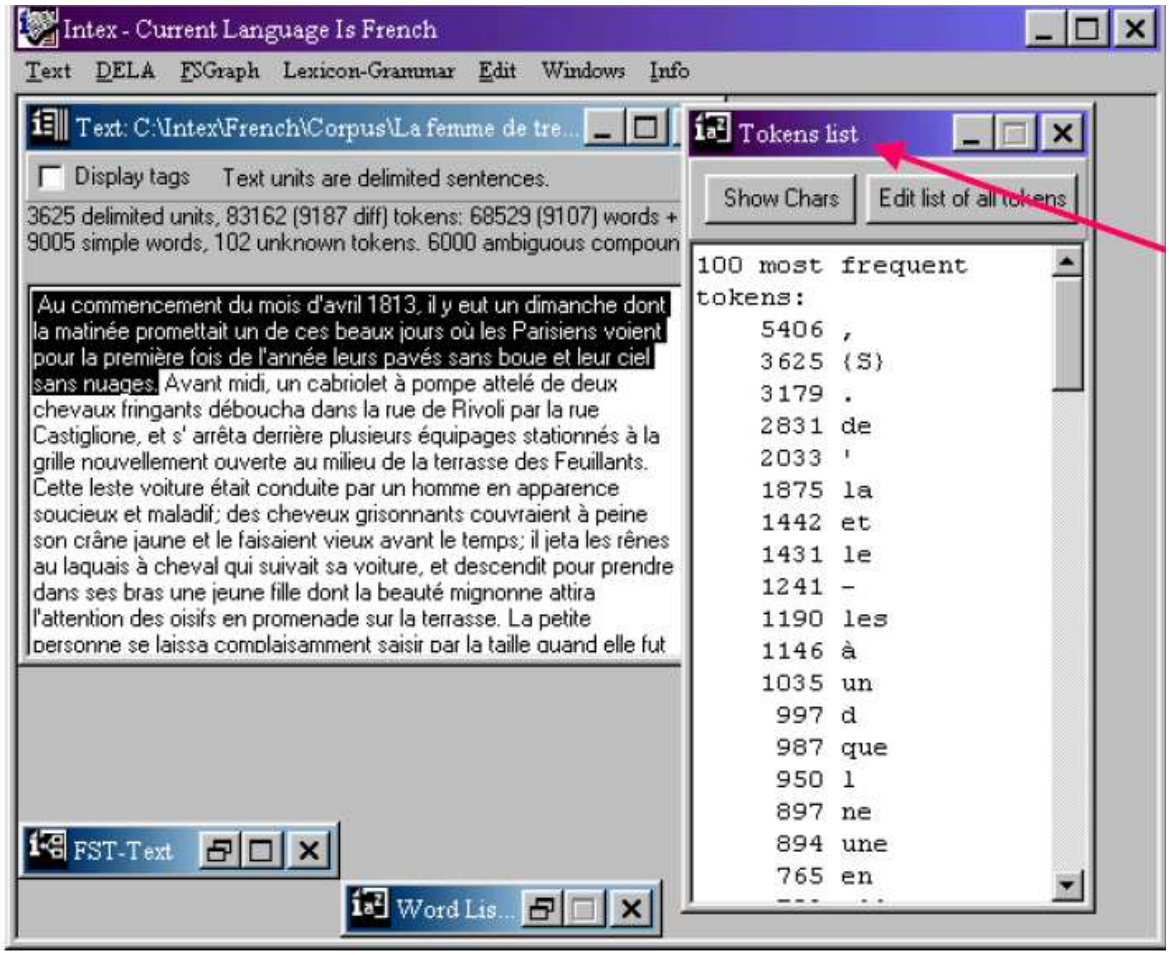


Figure.IV.4. *Les informations statistique du texte*

A partir de ces informations, il est donc possible de voir le nombre de phrases qui composent le texte et également celui des *tokens*» . Il existe quatre sortes de tokens qui sont en fait des objets de base de l'analyse par Intex. Les tokens peuvent représenter les *formes simples* qui apparaissent dans le texte. Les formes simples sont à distinguer des mots simples. Les formes simples sont des séquences de lettres enclavées par deux délimiteurs. Des *tags* qui sont un autre type de tokens représentent des données linguistiques et sont notées entre deux crochets (« { », «} »). Les *digits*, troisième type de tokens, correspondent aux chiffres de 0 à 9. Les délimiteurs, le dernier type de tokens, représentent des caractères autres qu'une lettre, un chiffre ou un espace. Ces informations statistiques sont archivées dans un ficher nommé « result.rtf » qui figure dans le répertoire courant.

> *Les lexèmes (tokens) sont les objets atomiques INTEX, classés en quatre types :*
> *-- les formes simples (simple forms) sont des séquences de lettres entre deux séparateurs ;*

-- les étiquettes (tags) représentent des informations linguistiques, et sont écrites entre accolades « { » et « } » ;

-- les chiffres (digits) sont les chiffres arabes (les dix caractères « 0 » à « 9 ») ;

*-- les séparateurs sont tous les caractères qui ne sont ni des lettres, ni des chiffres, ni des blancs. La séquence « **…** » par exemple est constituée de trois séparateurs. Par ailleurs, les blancs sont des séquences constituées des trois caractères :l'espace, le caractère de tabulation et le changement de ligne/paragraphe .*[1]

Outre ces informations statistiques, il est aussi possible d'obtenir des informations d'ordre linguistique sur la nature des formes qui composent le texte. Intex reconnaît en effet quatre types d'unités lexicales : les affixes qui sont des morphèmes dérivationnels ou flexionnels (préfixes ou suffixes pour le français), les mots simples, les mots composés (autrement dit les séquences constituées de plusieurs mots simples) et des expressions figées. Dans le cadre où apparaissent ces informations figure aussi une indication sur la notion d'ambiguïté. Une forme sera effectivement considérée comme étant ambiguë quand deux entrées des dictionnaires DELA correspondent à cette même forme. La notion de token permet justement d'éclaircir ce point. Le token correspond à la forme graphique que prend un mot simple qui figure dans le dictionnaire DELAS. Au mot simple «Le » correspondent les trois tokens suivants : « le », « Le », et « LES ». Mais seule une entrée de dictionnaire représentent ces trois tokens à savoir l'entrée « le , déterminat». Cette distinction entre les différentes unités d'analyse d'Intex est importante dans la mesure où les transducteurs ont recours à ces données du texte.

INTEX permet de traiter quatre type d'unités linguistiques :

*-- **les morphèmes** (préfixes, affixes, suffixes) sont des séquences de lettres inclues dans des formes simples, et associées à des informations linguistiques dans des graphes (morphologiques) ;*

*-- **les mots simples** sont des formes simples qui correspondent à une ou plusieurs entrées lexicales dans un dictionnaire (de mots simples). Les formes simples qui ne correspondent à aucune entrée lexicale sont des formes simples inconnues ;*

*-- **les mots composés** sont des séquences de formes simples qui correspondent à une*

ou plusieurs entrées lexicales dans un dictionnaire (de mots composés) ;

[1] Ibid, p.19

-les expressions figées *sont des séquences éventuellement discontinues de formes simples qui correspondent à des entrées lexicales dans une grammaire lexicale (d'expressions figées).*[1]

3.2.2.2. La recherche d'un motif dans le texte

Une fois que le texte a été chargé, il est possible de procéder à une recherche dans ce texte. La fenêtre qui apparaît en cliquant sur la fonctionnalité « Locate Pattern… » (Menu Text) permet de prendre en compte plusieurs paramètres qui permettent d'affiner la recherche.

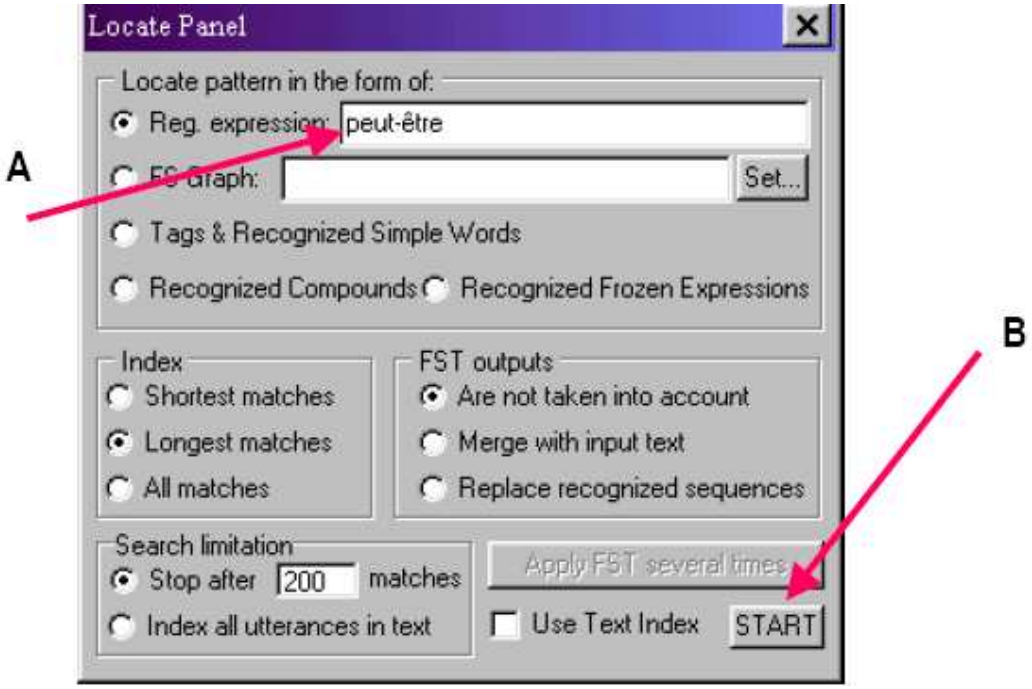


Figure.IV.5. *Boite de dialogue pour la recherche du motif peut-être*

Le motif de recherche apparaît sous la forme de lien hypertexte dans la fenêtre contenant le texte une fois que la recherche a été lancée. Une boîte de dialogue apparaît également à la fin de la recherche : cette boîte de dialogue qui porte l'entête « Display indexed sequences » permet de construire une concordance du motif de recherche. La construction d'une concordance pour une séquence donnée consiste à élaborer une liste des occurrences de cette séquence en affichant également son contexte.

[1] Ibid, p.19

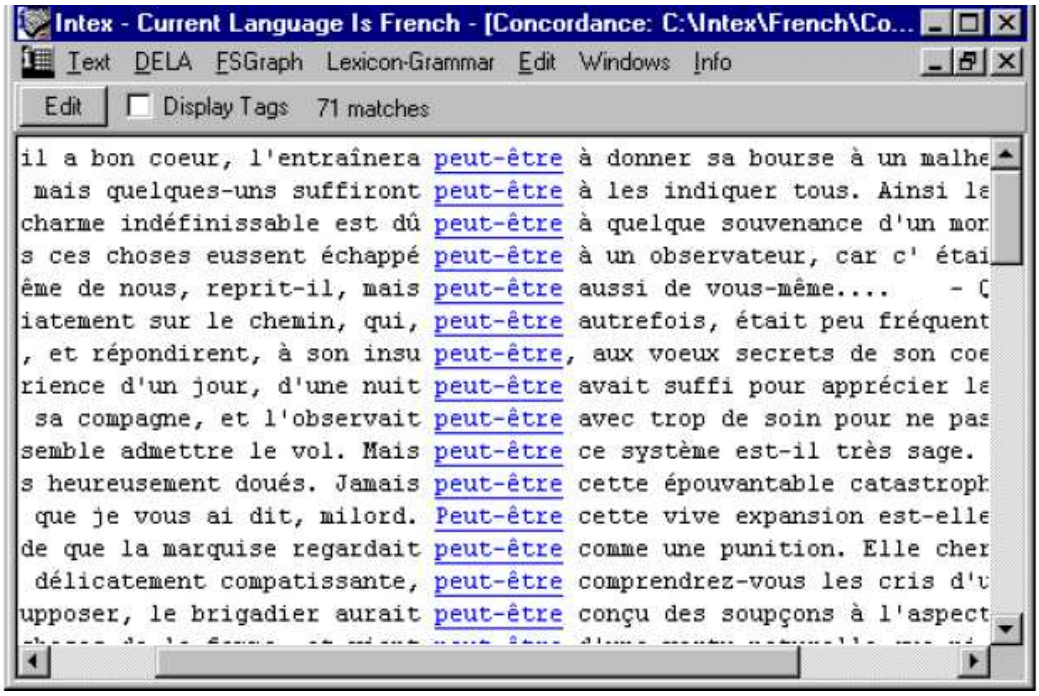


Figure.IV.6. La liste des occurrences ou *la concordance* du mot peut-être

Intex nous offre, aussi la possibilité de paramétrer le nombre d'éléments pouvant apparaître dans le contexte gauche ou le contexte droit du motif recherché en plus de diverses options qui permettent notamment de rechercher les formes fléchies d'un mot avec un unique motif de recherche. Il serait donc possible de filtrer toutes les formes fléchies d'un verbe à partir d'un unique motif de recherche ce qui peut s'avérer fort efficace dans le cas des verbes du troisième groupe qui présentent de nombreuses irrégularité

3.2.2.3. Application des transducteurs

Une autre fonctionnalité d'Intex , tout aussi utile consiste dans le traitement de corpus. La phase de traitement est précédée par une phase de préparation du texte qui s'opère par l'intermédiaire d'une fenêtre qui porte l'entête « Preprocessing a Text » lors du chargement du texte. Cette fenêtre permet d'appliquer les transducteurs correspondant à la langue du texte et les diverses ressources linguistiques telles que les dictionnaires électroniques DELA

Un des transducteurs appliqués au texte permet de segmenter le texte. Il s'agit du transducteur « sentence.fst » qui permet de remplacer les délimiteurs superflus en d'insére des caractères tels que « {S} » qui représentent un retour à la ligne.

3.2. 2.4. Application des ressources lexicales

La fonction qui nous intéresse ici c'est celle que nous trouvons dans l'onglet « Apply lexical Resources » dans le menu « Text » permet de procéder à une analyse

lexicale en appliquant donc les ressources lexicales disponibles dans l'application. La boite de dialogue qui apparaît comporte trois zones distinctes dédiées respectivement aux mots simples, aux mots composés, et aux expressions figées frozon expresions. Dans ces zones apparaissent donc les outils, dictionnaires et transducteurs, permettant de procéder à une analyse morphologique. Les dictionnaires électroniques DELAF ou DELACF sont ceux qui sont utilisés pour cette opération

La zone qui nous intéresse le plus est celle qui concerne les expressions figées. Les transducteurs lexicaux qui figurent dans cette zone permettent de représenter chaque expression figée sous forme de graphe. Ce graphe désigne une expression figée dans sa structure de base ainsi que toutes les variantes qu'elle connaît.. Les tables répertoriées Cxxx correspondent au lexique-grammaire des expressions figées : les noms de fichiers qui ont pour nom Cxxx correspondent donc aux noms des tables du lexique-grammaire. Intex a notamment recours à la table du lexique-grammaire notée C1d.xsl.

3.2. 2.5. Exemple de transducteur lexical d'expressions figées

Ce transducteur lexical, reconnaît les variantes de l'expression *Perdre la raison,* qui peuvent éventuellement contenir une insertion « *Luc a* perdu *soudain* les pédales »

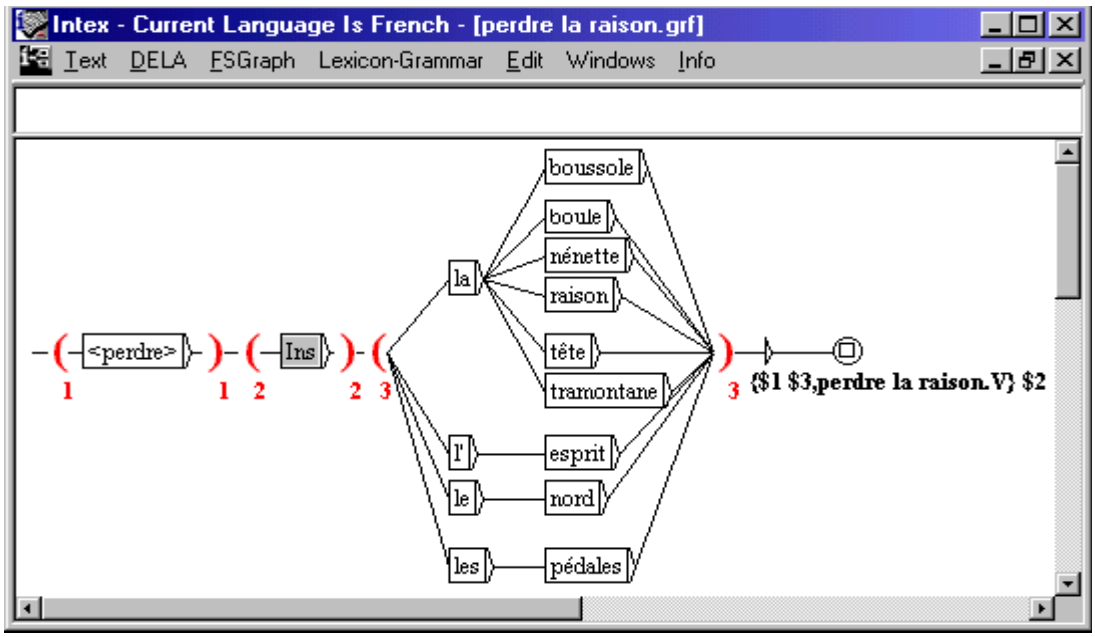


Figure.IV.7. *Un graphe d'une expression figée e ses variantes*

En effet ces transducteurs tiennent compte d'insertions qui peuvent apparaître entre les constituants de l'expression et les « faire sortir » des entrées lexicales (rappelons que cette
caractéristique distingue par définition les expressions figées des mots composés).
Par exemple, si le transducteur précédent est appliqué au texte : « *Luc a perdu soudain les pédales* », la forme « *perdu* » va être rangé dans la variable **$1**, la forme « *soudain* » dans la variable **$2**, et la séquence « *les pédales* » dans la variable **$3**. Le résultat produit est alors :
Luc a {perdu les pédales, perdre la raison. V} soudain

3.2.3. Les tables de lexique grammaire dans l'application Inetx

En premier lieu, les tables de lexique-grammaire des expressions figées ont été construites à la main. Puis le LADL a recensé pour le français plus de 30 000 expressions figées, rangées dans les tables **Cxx** du lexique-grammaire . Les expressions figées qui sont rangées dans une même table partagent la même construction syntaxique de base. INTEX permet d'automatiser la construction du transducteur pour chaque table. Pour cela, il faut :
(1) une table de lexique-grammaire ;
(2) un graphe-patron qui formalise les propriétés décrites dans la table. [1]

INTEX peut alors mettre en correspondance les propriétés de chaque entrée de la table et les chemins du graphe-patron correspondant ; le résultat est un transducteur qui a la même fonction et la même forme que le transducteur que l'on aurait construit à la main, mais qui peut représenter plusieurs centaines d'expressions.

Pour pouvoir être utilisée par INTEX, la table du lexique-grammaire doit bien entendu être entrée en machine. La façon la plus naturelle est d'entrer la table avec un tableur, comme par exemple Microsoft Excel : chaque entrée est décrite sur une ligne ; les propriétés sont entrées en colonne ; une cellule du tableur (à l'intersection d'une ligne et d'une colonne) contient soit du texte, soit un signe « + » ou « - ».Le format d'une table de lexique-grammaire doit respecter quelques contraintes, naturelles pour tous ceux qui connaissent le lexique-grammaire :

1. La première ligne du tableau contient le nom de chaque zone de texte ou propriété le «titre » de chaque colonne) ; il doit y avoir autant de titres que de colonnes ;

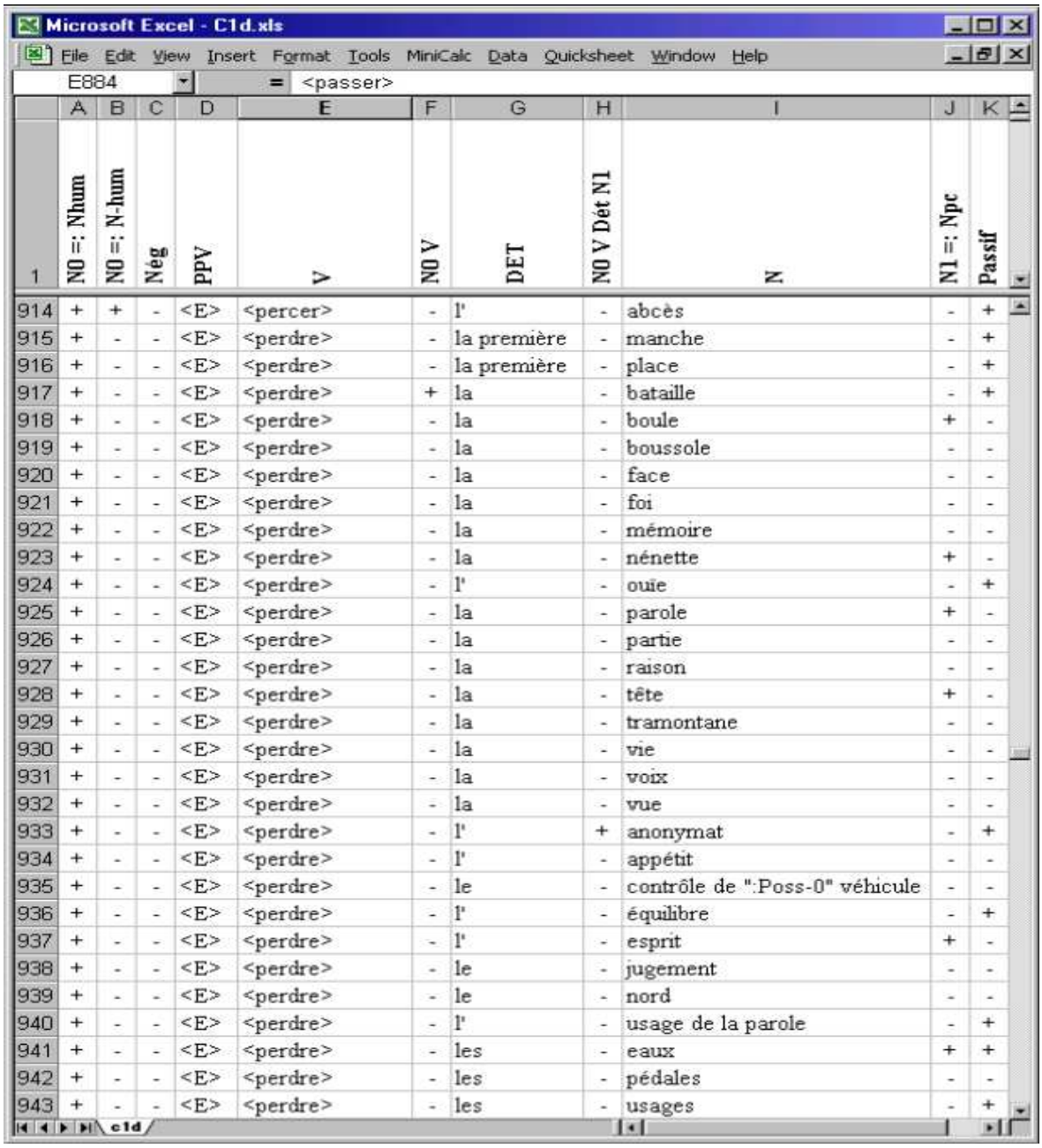
[1] Ibid, p. 163

2. les entrées du tableau commencent à la seconde ligne ; chaque entrée du lexiquegrammaire est décrite sur une ligne au maximum ; il ne doit pas y avoir de ligne vide ;

3. une cellule de texte contient une séquence qui peut comporter des constantes (ex. « la raison »), des symboles lexicaux (ex. « <perdre> », « <aller:P> » ou « <PREP> ») et des références à des graphes (ex. « :Dnum ») ; le symbole « <E> » doit être utilisé pour représenter la séquence vide ;

4. une cellule de propriété contient exclusivement un caractère « + » ou « - » ; les colonnes doivent être homogènes : une colonne de propriété ne peut contenir aucun texte ; une colonne de texte ne peut pas contenir de signe « + » ou « - ».

Pour le bien illustrer nous présentons ci-dessous un extrait de la table C1d des expressions figées



Microsoft Excel - C1d.xls

File Edit View Insert Format Tools MiniCalc Data Quicksheet Window Help

E884 = <passer>

	A	B	C	D	E	F	G	H	I	J	K
1	N0 =: Nhum	N0 =: N-hum	Nég	PPV	V	N0 V	DET	N0 V Dét N1	N	N1 =: Npc	Passif
914	+	+	-	<E>	<percer>	-	l'	-	abcès	-	+
915	+	-	-	<E>	<perdre>	-	la première	-	manche	-	+
916	+	-	-	<E>	<perdre>	-	la première	-	place	-	+
917	+	-	-	<E>	<perdre>	+	la	-	bataille	-	+
918	+	-	-	<E>	<perdre>	-	la	-	boule	+	-
919	+	-	-	<E>	<perdre>	-	la	-	boussole	-	-
920	+	-	-	<E>	<perdre>	-	la	-	face	-	-
921	+	-	-	<E>	<perdre>	-	la	-	foi	-	-
922	+	-	-	<E>	<perdre>	-	la	-	mémoire	-	-
923	+	-	-	<E>	<perdre>	-	la	-	nénette	+	-
924	+	-	-	<E>	<perdre>	-	l'	-	ouïe	-	+
925	+	-	-	<E>	<perdre>	-	la	-	parole	+	-
926	+	-	-	<E>	<perdre>	-	la	-	partie	-	-
927	+	-	-	<E>	<perdre>	-	la	-	raison	-	-
928	+	-	-	<E>	<perdre>	-	la	-	tête	+	-
929	+	-	-	<E>	<perdre>	-	la	-	tramontane	-	-
930	+	-	-	<E>	<perdre>	-	la	-	vie	-	-
931	+	-	-	<E>	<perdre>	-	la	-	voix	-	-
932	+	-	-	<E>	<perdre>	-	la	-	vue	-	-
933	+	-	-	<E>	<perdre>	-	l'	+	anonymat	-	+
934	+	-	-	<E>	<perdre>	-	l'	-	appétit	-	-
935	+	-	-	<E>	<perdre>	-	le	-	contrôle de ":Poss-0" véhicule	-	-
936	+	-	-	<E>	<perdre>	-	l'	-	équilibre	-	+
937	+	-	-	<E>	<perdre>	-	l'	-	esprit	+	-
938	+	-	-	<E>	<perdre>	-	le	-	jugement	-	-
939	+	-	-	<E>	<perdre>	-	le	-	nord	-	-
940	+	-	-	<E>	<perdre>	-	l'	-	usage de la parole	-	+
941	+	-	-	<E>	<perdre>	-	les	-	eaux	+	+
942	+	-	-	<E>	<perdre>	-	les	-	pédales	-	-
943	+	-	-	<E>	<perdre>	-	les	-	usages	-	+

c1d

Figure.IV.8.*Extrait de la table C1d*[1]

Cette extrait contient toutes les expressions figées de la table **C1d** qui contiennent le verbe perdre. La table **C1d** décrit plus de 1 500 expressions figées dont la structure syntaxique est : **N0 V N1**, où **N0** représente le sujet, **V** le verbe et **N1** le complément d'objet direct figé.

[1]Ibid, p. 165

Pour utiliser ces tables afin d'extraire des expressions figées d'un corpus l'application Intex fait appel à d'autres programmes représentés sous forme de graphes qui les aident à résoudre certaines difficultés à savoir :

-Différencier les formes constants ou ce qu'on appelle la zone fixe (*ex : raison*) des formes qui peuvent être fléchies (*ex : perdre*)

- Reconnaissance des formes les formes : *à, au, aux, de, les, du le, la , les, l'.*

- quelques formes de conjugaison au temps composés :

(se casser la figure) Luc et Marie **se** *sont* **cassé la figure**

(en faire le moins possible) Paul **en** *a toujours* **fait le moins possible.**

De ce fait l'application fait appel à des min graphes pour prendre en charge ces difficultés ; ainsi le graphe *poss* reconnaît *les formes* mon, ton, son, ; le graphe *à* reconnaît les formes à, au, aux etc. .

Par ailleurs tout table de lexique-grammaire doit avoir un graphe patron pour la décrire et l'interpréter ;la table C1d des expressions figées à un graphe-patron qui permet leur reconnaissance dans un corpus et voila un graphe-patron simplifiés de la table c1d .

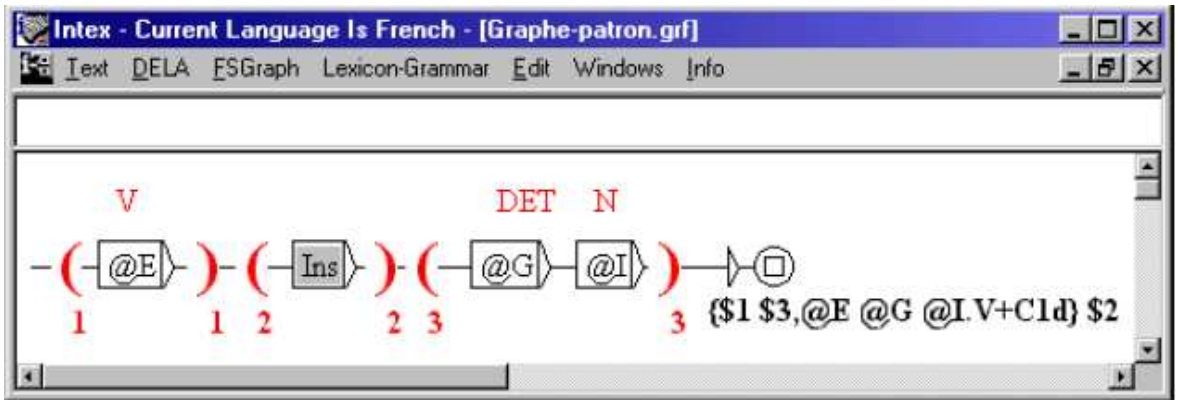


Figure.IV.9.*Un graphe-patron de la table C1d*

Les variables E, G, I sont présentes en *entrée* du transducteur (graphe) pour identifier les constituants des expressions dans le texte, et aussi en *sortie* du transducteur pour lemmatiser les expressions reconnues. Par exemple, l'étiquette produite par le transducteur précédent contiendra comme lemme le texte présent dans les cellules **E**, **G** et **I** de la table.

3.3. Traitement du corpus au moyen de l'application Intex

Nous avons tout d'abord télécharger le logiciel Intex que nous avons choisi pour le traitement de notre corpus du cite http://intex.univ-fcomte.fr/

3.3.1. les informations statistiques du corpus

Nous chargeons notre corpus que nous avons déjà préparé en le conversant au format approprié à ce logiciel ; nous utilisons le menu Text > Open et nous ouvrons le fichier qui contient notre corpus constitué de quelques extraits des différents numéros du journal *El Watan* du mois de janvier 2007 ainsi que des suppléments *économi*e et *immobilie*r ,dépendant du même journal . Voila le résultat donné par Intex

Ce premier traitement nous fournit les renseignements affichés dans l'écran suivant :

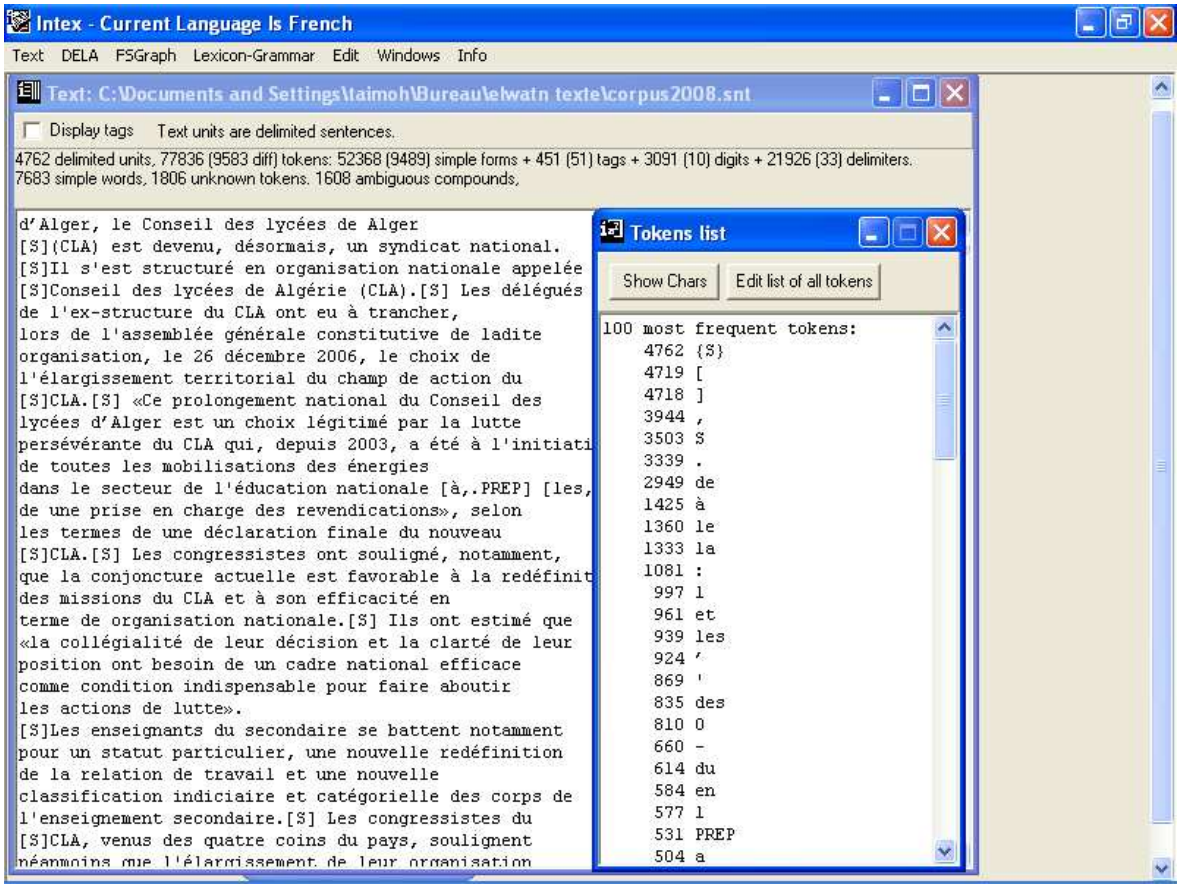


Figure.IV.10. *Informations statistiques du corpus et fréquence des mots*

Ce premier écran nous montre que :

-Le corpus se compose de 4762 unité textuelle ou **phrases** (*delimited units*) ;ce découpage est faits automatiquement selon les normes spécifiques de cette application

-77836 **lexèmes** (*tokens)* qui représentent les objets de base de l'analyse pour I ntex et dont 9583 (différents).Intex distingue trios types de lexèmes :

-52368 **formes simples** dont 9489(différents)

-451 **tags** dont51 (différents) :(notes ou commentaires centenants des infomations linguistique et concerne les dictionnaires et les tables de lexique-grammiare)

-3091 **Chiffre**s (digits) dont 10 différents.

-21926 **séparateurs** (dilimiters)dont (33) différentes ; ce qui représente tout les symboles et les marques de ponctuation .

- 7683 **mot simple** (simple words) qui sont à distinguer des formes simple mots

-1806 **lexèmes inconnus** (unknowon tokens) .

-1608 **formes ambiguës** (ambiguous compounds) .

Intex nous affiche aussi, dans ce premier résultats les 100 lexèmes les plus fréquents dans le corpus ; nous signalons ici que le terme lexème (token) comprend aussi les symboles et les marques de ponctuation .

Nous observons dans cette liste que le mots le plus fréquent est la forme « *de* » puis « *la* » ,puis d'autres morphèmes syntaxiques sans compter le *« {s}* » qui signifie les blancs et les marques de ponctuation, chiffre et articulateurs et verbes auxiliaires et modales

Cette application nous fournis les 100 formes les plus fréquents seulement nous des résultats plus exhaustifs avec une application ou la fréquence des mots pourrait nous donner des indications sur la thématique du corpus .

3.3.2. Analyse lexicale

3.3.2.1. Préparation des ressources lexicales

Pour faire l'analyse lexicale du corpus, le programme doit appliquer ce quel' on appelle les ressource lexicales que nous avons défini dans une section précédente, notamment les dictionnaires électronique et les tables de lexique-grammire

L'écran ci-dessous présentes les ressource lexicales disponibles sur cette application ; ils sont partagées en trois types :

l-es dictionnaire des mots simples, dont le dictionnaire est Delaf.bin

2-les dictionnaires des mots composés dont nous avons choisi Delacf.bin pour notre analyse puisque il s'agit d'une base de données qui permet de reconnaître 248.885 mots composé dont : -236.969 noms composé

-4732 locutions adverbiales.

-4356 locutions prépositives

-2119 locutions adjectivales.

-615 phrases nominales.

-50 pronoms composé.

-44 locutions conjonctives

3- les tables des expressions figées (*frozen expressions*) qui sont ici distingués des mots composé puisque ces séquence pourraient être discontinu c'est-à-dire on peut y insère des éléments libres et ils peuvent subir des variation flexionnelle ; par exemple : *Perdre la raison ; il a perdu* **soudain** *la raison* . En appliquant la table de lexique-grammaire *Cd1.cfg* .cette application peut reconnaître 1663 locutions verbales et prendre en compte toutes les variantes mentionnées sur la table.

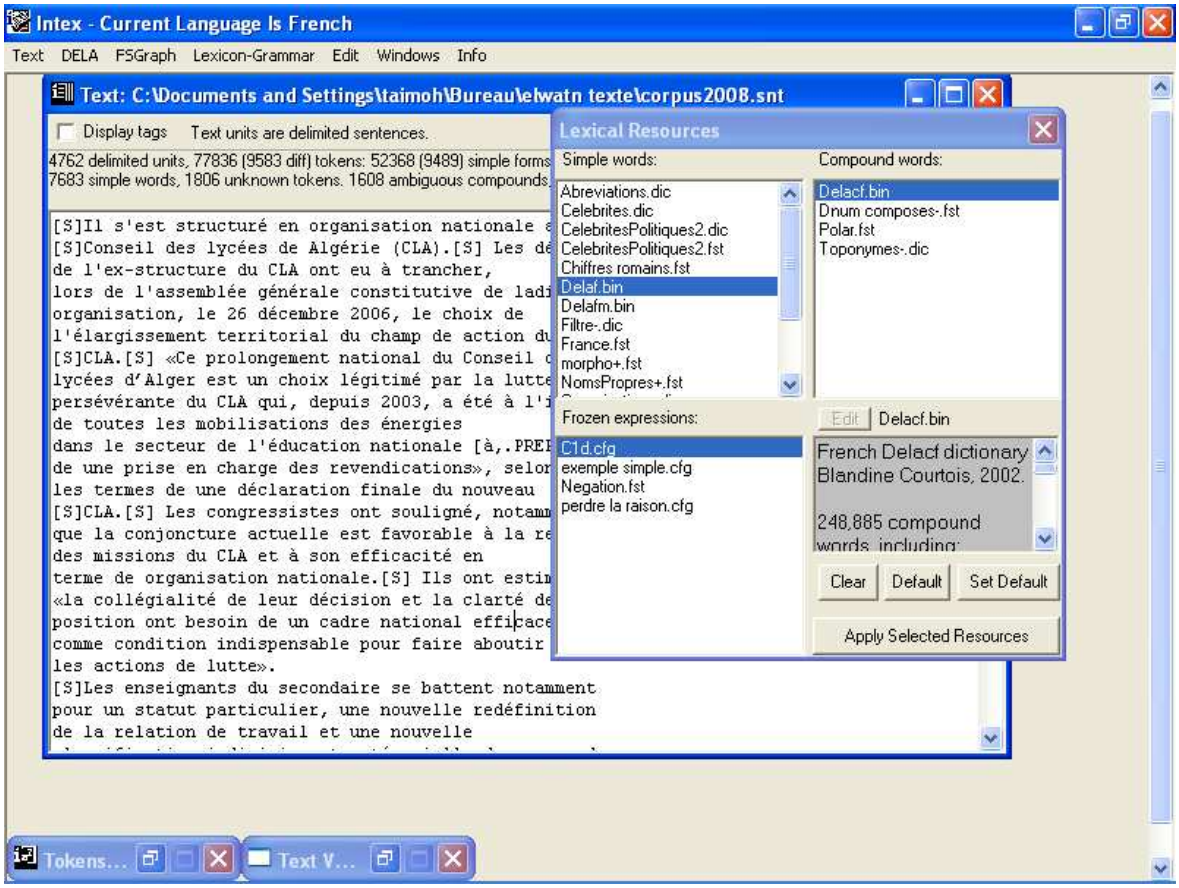


Figure.IV.11.*Les ressources lexicales disponibles sur l'application Intex*

3.3.2.2. Application des ressources lexicales

Les ressources mentionnées en bleue sont les ressources choisies pour notre recherche. Nous lançons l'analyse en cliquant sur l'onglet *Apply Selected Resources* et le logiciel va appliquer les transducteurs correspondant à ces dictionnaire notamment le graphe de la table Cd1 qui nous intéresse le plus .Après quelque minutes s'affiche le résultat suivant :

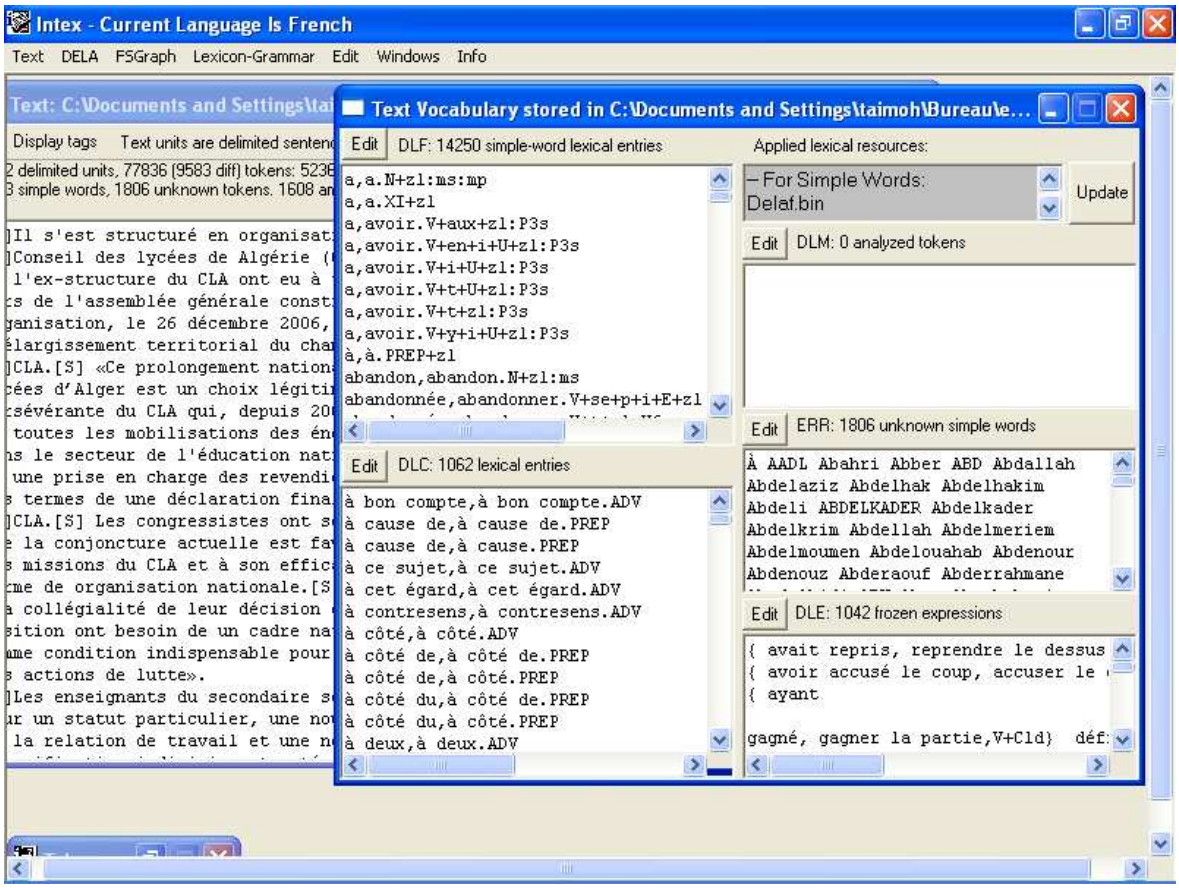


Figure.IV.12. *Résultat de l'analyse lexicale*

Nous observons dans la boite de dialogue supérieur quatre listes à savoir la liste des14250 mots simples , la liste des1806 mots simples non reconnus (*unknown simple words*) , la liste des1062 mots composés (*DLC)* et la listes de 1042expressions figées (frozen expressions) reconnues à partir de la tables C1d.

3.3.3. Edition des résultats requis

Notre recherche s'intéresse aux deux dernières listes ; la première qui comprends des nom composés et différents types de locutions et la deuxième qui contient les locutions verbales employés dans le corpus. Pour extraire les résultats du logiciel nous faisons leur édition l'édition en cliquons sur l'onglet *edit* en haut de la liste. Après la conversion des fichier obtenus au format *word* voila un extrait des deux listes :

3.3.3.1. le résultat pour les expression figées (*frozen expressions*)

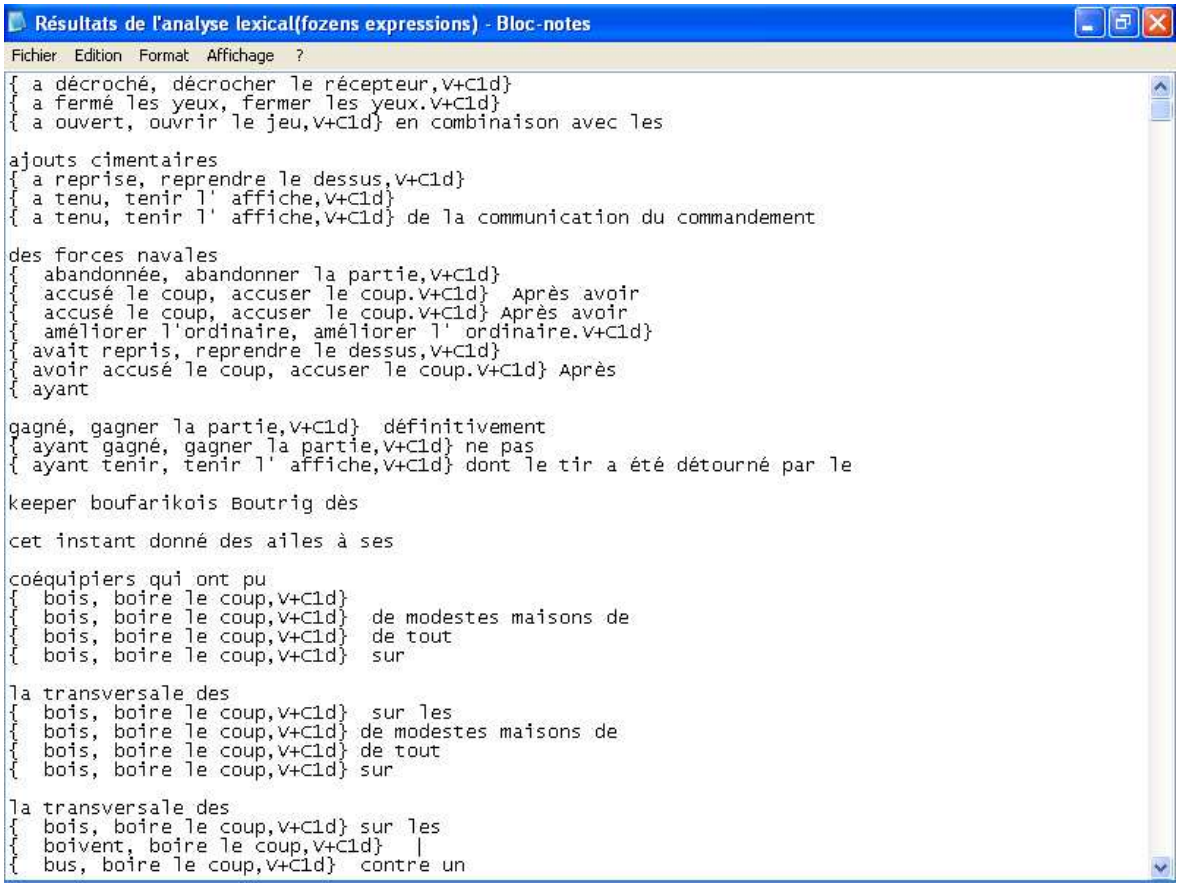


Figure.IV.13. *Extraits de la liste des expressions figées (locutions verbales) produite par Intex*

Cette liste représente des occurrences des locutions verbales : *décroche le récepteur, fermer les yeux, ouvrir le jeu, reprendre le dessus, , tenir l'affiche, abandonner la partie, accuser le coup,améliorer l'ordinaire, gagner la partie* . Et la liste compte 1042 locutions .

Nous observons dans cette liste que ce programme attribue à expression son lemme (sa forme canonique) ex : *a décroché l'appareil* son lemme c'est décrocher l'appareil. Cette opération on l'appelle en TAL la *lemmatisation* [1]. Ensuite, il reconnaît la distribution syntaxique de l'expression (*v+cd1* veut dire verbe+ compliment d'objet direct) ; cette deuxième étape s'appelle *l'étiquetage*[2]

[1] La lemmatisation consiste à remplacer une forme fléchie par son lemme. Le lemme constitue la forme de base d'un mot donné. La lemmatisation présente donc cette forme de base sans aucune marque de flexion (pluriel, désinence ou forme conjuguée d'un verbe). La lemmatisation permet de remplacer une forme actualisée par sa forme canonique. Le lemme ou la forme canonique d'un mot constitue une entrée de dictionnaire.

[2] l'étiquetage constitue une étape importante dans le traitement automatique de corpus. Il 'agit d'une opération de base qui vise à étiqueter les formes pertinentes d'un texte ayant le statut d'unités de base.

3.3.3.2 les résultat des mots composés *(compound words)*

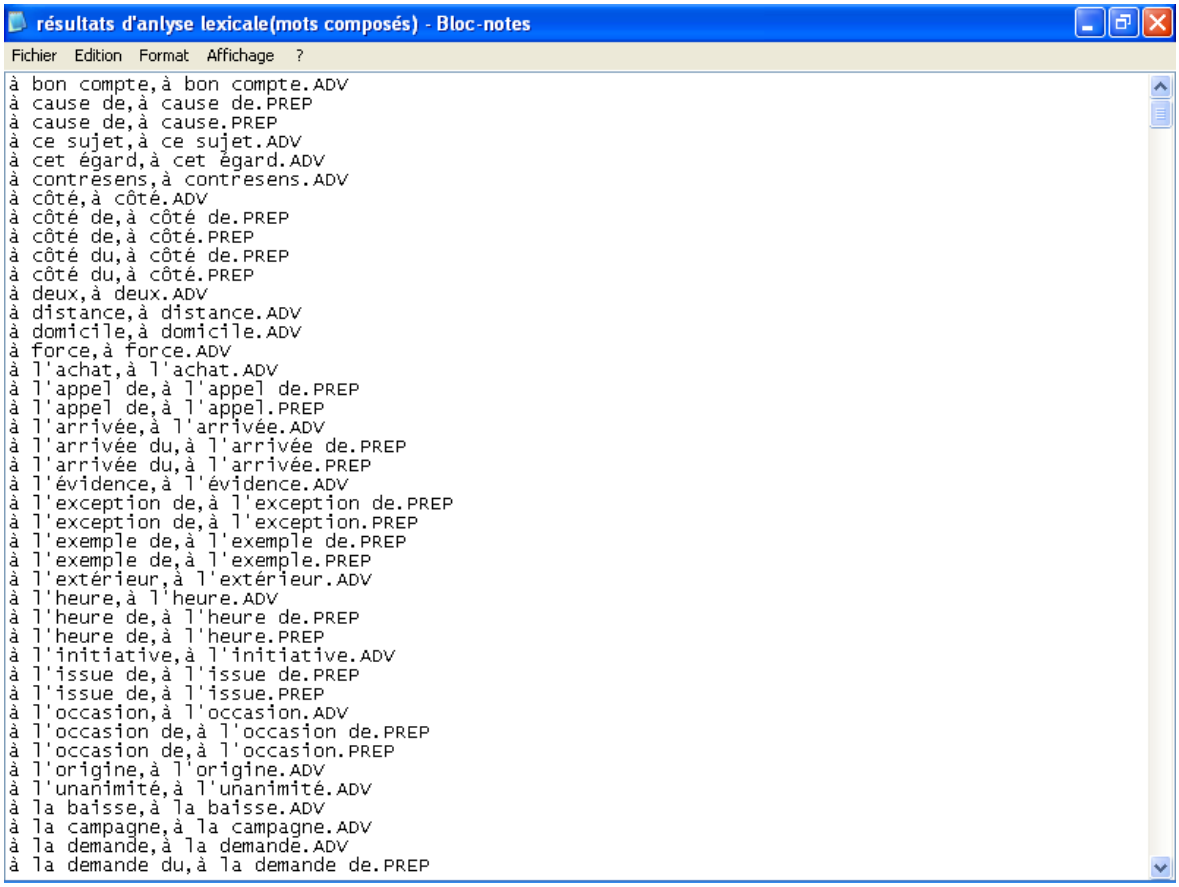


Figure.IV.14. *Extrait de la liste des mot composé (compound words) produite par Intex*

3.3.4. Analyse des résultats produits par Intex

Nous considérons les séquences des deux listes comme des expressions figées et nous allons essayer, par la suite d'en faire une étude statistiques et typologique

Pour ce faire, nous allons faire recourt à une autre logiciel. Il s'agit de l'application *Nooj* qui est dévloppée à partir de Intex par Max Selberstein au Laboratoire de Sémio-Linguistique et Didactique (LASELDI) de l'Université de Franche-Comté. *Nooj* est conçu comme environnement de développement linguistique qui permet de construire et de gérer des dictionnaires et grammaires électroniques à large couverture afin de formaliser divers niveaux des langues naturelles : orthographe, morphologie flexionnelle et dérivationnelle, lexique de mots simples, mots composés et expressions figées,

L'étiquetage morphosyntaxique consiste donc dans l'affectation automatique d'étiquettes morphosyntaxiques aux formes constituants ces unités en fonction du contexte.

Tout d'abord nous chargeons les fichiers texte contenant les résultats dans l'application au moyen du menu Fil> open > text et le logiceil nous affiche l'écran suivante :

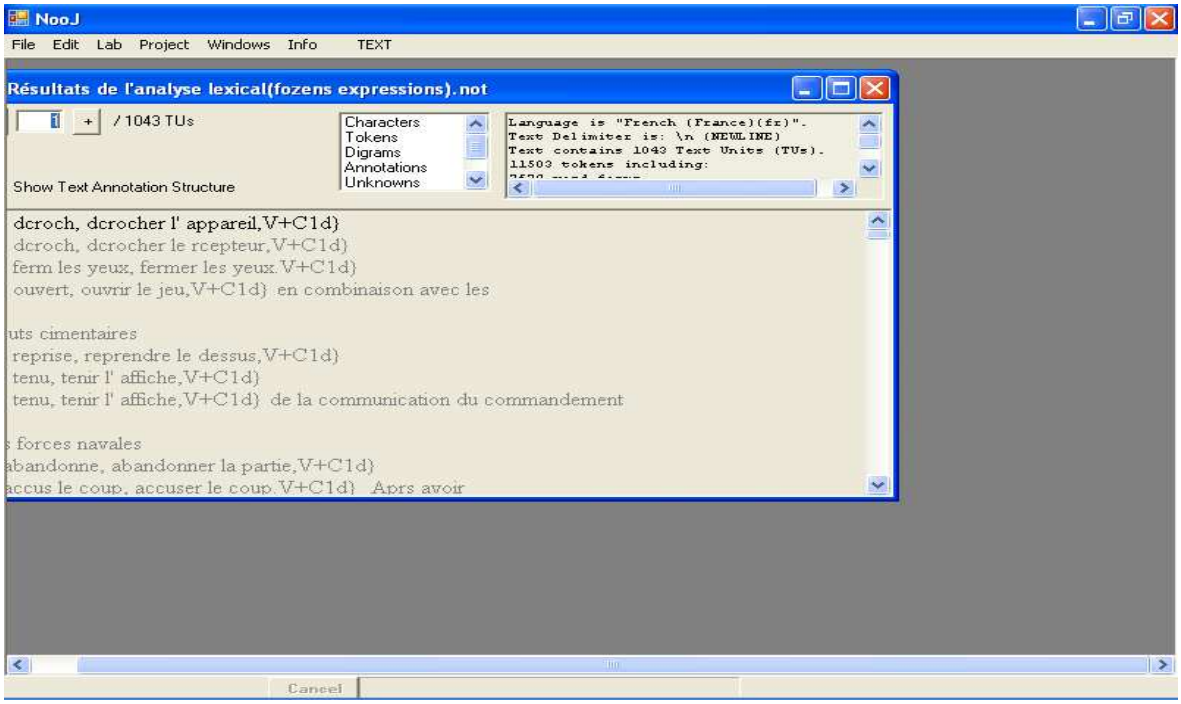


Figure.IV.15. *Analyse des résultats de Intex au moyen de l'aplication Nooj*

L'option *tokens* nous permet de reconnaître la fréquence des mots dans cette liste .C'est la fonction que nous allons exploiter pour compter la fréquence et les types des locutions verbales et de mots composés employés dans notre corpus .

3.3.4.1. Analyse des résultats de la partie *frozen espressions*

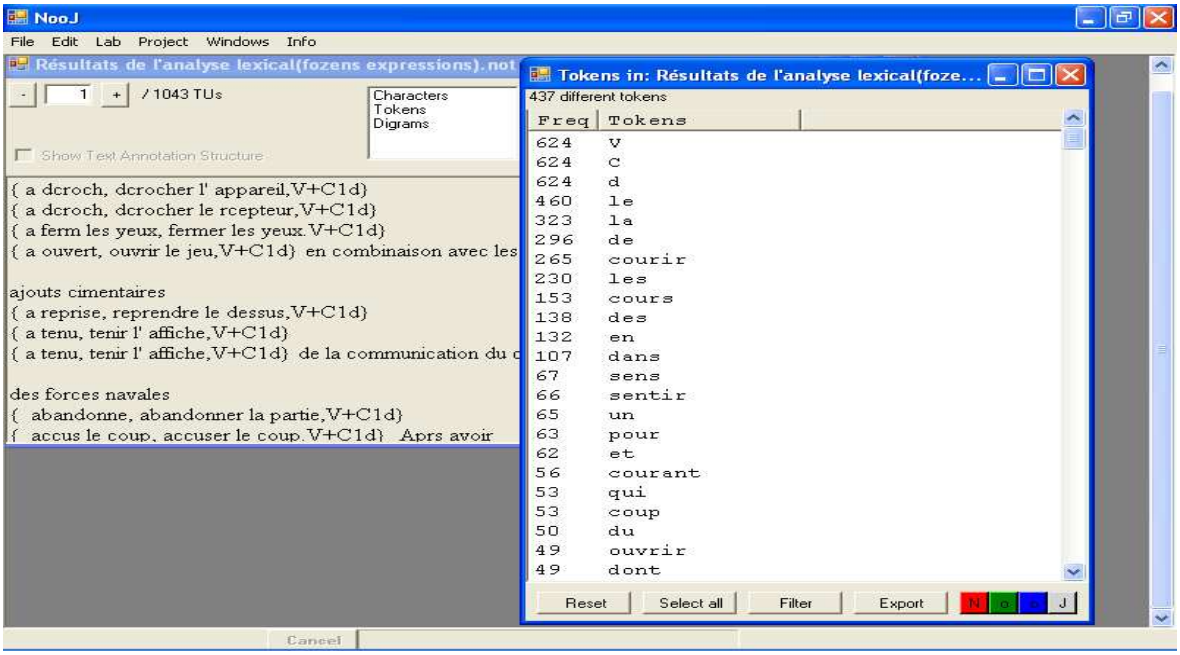


Figure.IV.16. *Fréquence des mots dans la liste des locutions verbales*

Nous observons que le mot *courir* à une fréquence de 339 ce que nous fait déduire que le corpus contient ce nombre de locutions à base verbale de *courir*. En cliquant sur l'onglet *Nooj* coloré en rouge en bas de l'écran , nous obtenons toutes les occurrences du verbe courir dans la liste des locutions et voila le résultat obtenu.

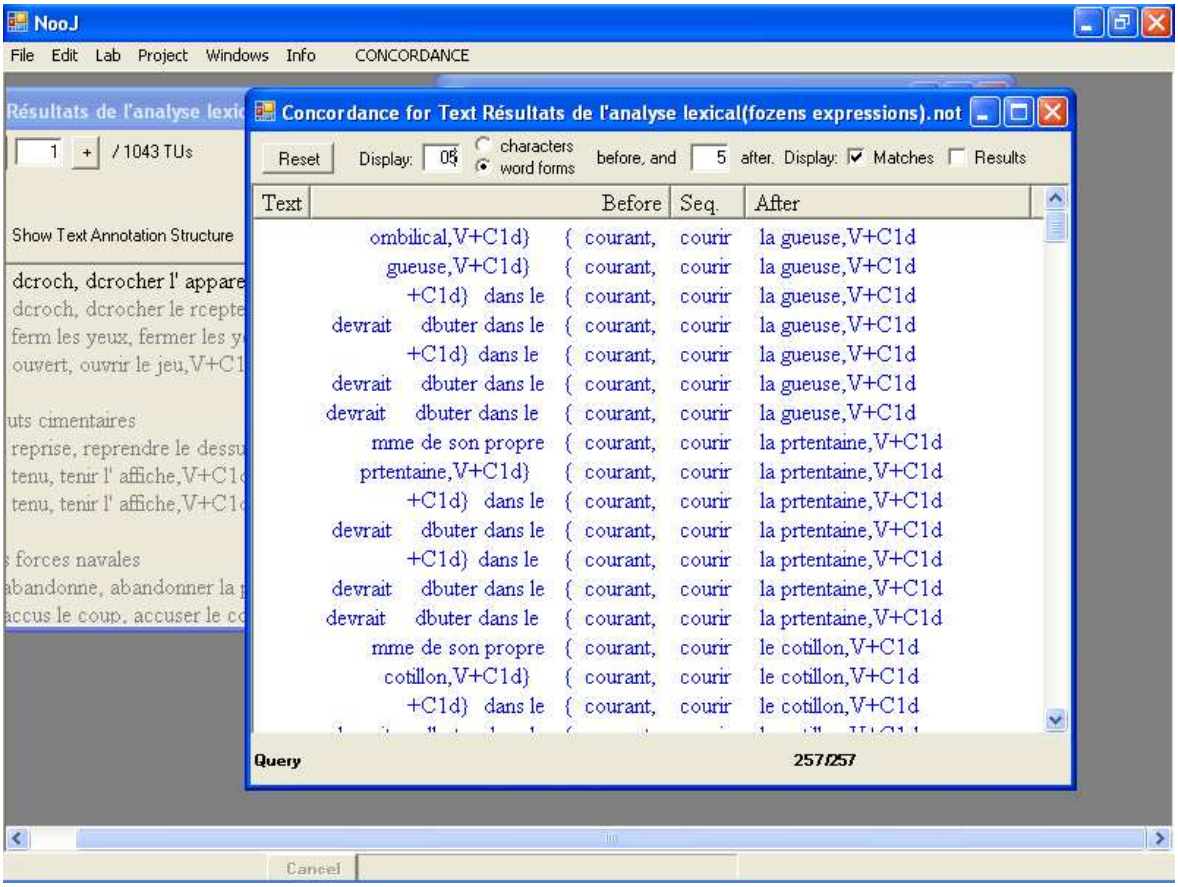


Figure.IV.17. *Les occurrences du verbe courir dans la liste des locutions verbales*

Le verbe suivant, c'est *sentir* (66 occurrences) puis *ouvrir* (51) etc. .

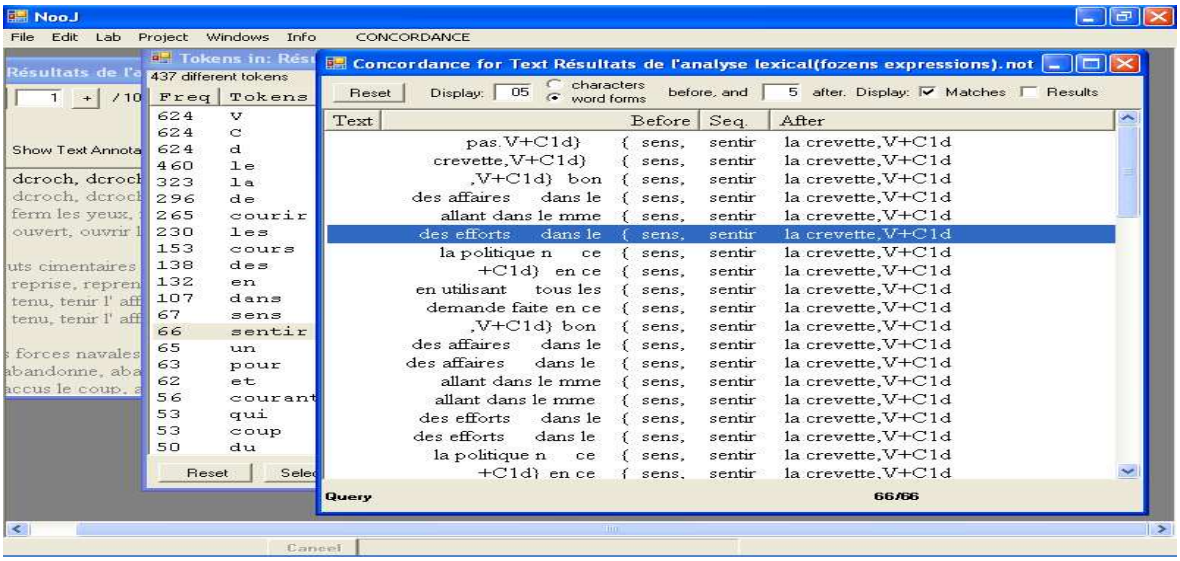


Figure.IV.18. *Les occurrences du verbes sentir employé dans des locutions verbales figées*

Pour bien illustrer les résultats de la recherche pour la première liste des locutions verbales figées, nous récapitulons l'analyse de ces résultats dans le tableau suivant :

Verbe de base	**Fréquence dans les locutions verbales de la table C1d**	**Locutions verbales figées**
1-courir.	288	*Courir la prétentaine.* *Courir les honneurs.* *Courir le guilledou.* *Courir le cotillon.* *Courir les filles.* *Courir Les Jupons……..*
2-sentir.	266	*Sentir la crevette.* *Sentir le fauve.* *Sentir le bouc.*
3-ouvrir.	47	*Ouvrir le jeu.* *Ouvrir le score.* *Ouvrir la marque.*
4-boir.	41	*Boire le coup.*
5-tenir.	39	*Tenir l'affiche*
6-reprendre.	21	*Reprendre le dessus*
7-chercher.	14	*Chercher La bagarre.*
8-calmer.	12	*Calmer les esprits.*
9-gagner.	12	*Gagner la partie.*
10-faire.	12	*Faire la bringue.* *Faire la différence.* *Faire le break*
11-remonter.	08	*Remonter la pente.*
12- perdre.	07	*Perdre la bataille.*
13- rendre.	06	*Rendre le dernier souffle.*
14-renversser	04	*Renverser la vapeur.*
15-resserer	04	*Resserrer la ceinture.*
16-bondonner.	02	*Abandonner la partie.*
17-décrocher.	02	*Décrocher l'appareil.*

18-creuser	02	*Creuser la caboche.* *Creuser la tête.* *Creuser les méninges.*
19- accuser.	02	*Accuser le coup.*
20-débarasser.	02	*Débarrasser la table.*
21-désservir.	02	*Desservir la table.*
22-mobiliser.	01	*Mobiliser les énergies*
23-défrayer.	01	*défrayer la chronique*
24-pater.	01	*pater la galerie*
25-remuer.	01	*Remuer la tête.*
26-trancher.	01	*Trancher le nœud gordien.*

Tableau.IV.4. locutions verbales figées détectées par l'application *Intex*

Après l'observation de ce tableau, et la vérification de la présence de ces locutions dans le corpus, nous nous sommes rendus compte que ces résultats ne sont pas fiables .Il s'est révélé que l'application notamment la fonctionnalité *frozen expresions* n'a pas utilisé la ta table de lexique-grammair C1d des locutions verbales correctement ; on a affecté à chaque forme verbale fléchée dans le texte des lemmes des locutions qui contiennent ce verbe . Il s'agit en fait d'un problème de lemmatisation qui dépend du fonctionnement du transducteur responsable de l'application de la table C1d.

3.3.5.2. Analyse des résultats de la partie *compound words*

Nous analysons , en second lieu , la deuxième liste de mots composés obtenus suite à l'application du dictionnaire *DELACF* dans la partie *compound words* de la fonction *apply lexical* ressources du logiciel *Intex*. Pour ce faire nous utilisons la même application *Nooj* , notamment la fonction *Tokens* pour déterminer la fréquence des mots dans cette liste, ce qui nous permettra de compter la fréquence des locutions et noms composés reconnus à partir de cette liste.

Après le chargement de cette liste dans le logiciel, et l'application de la fonction Tokens , nous aurons le résultat suivant :

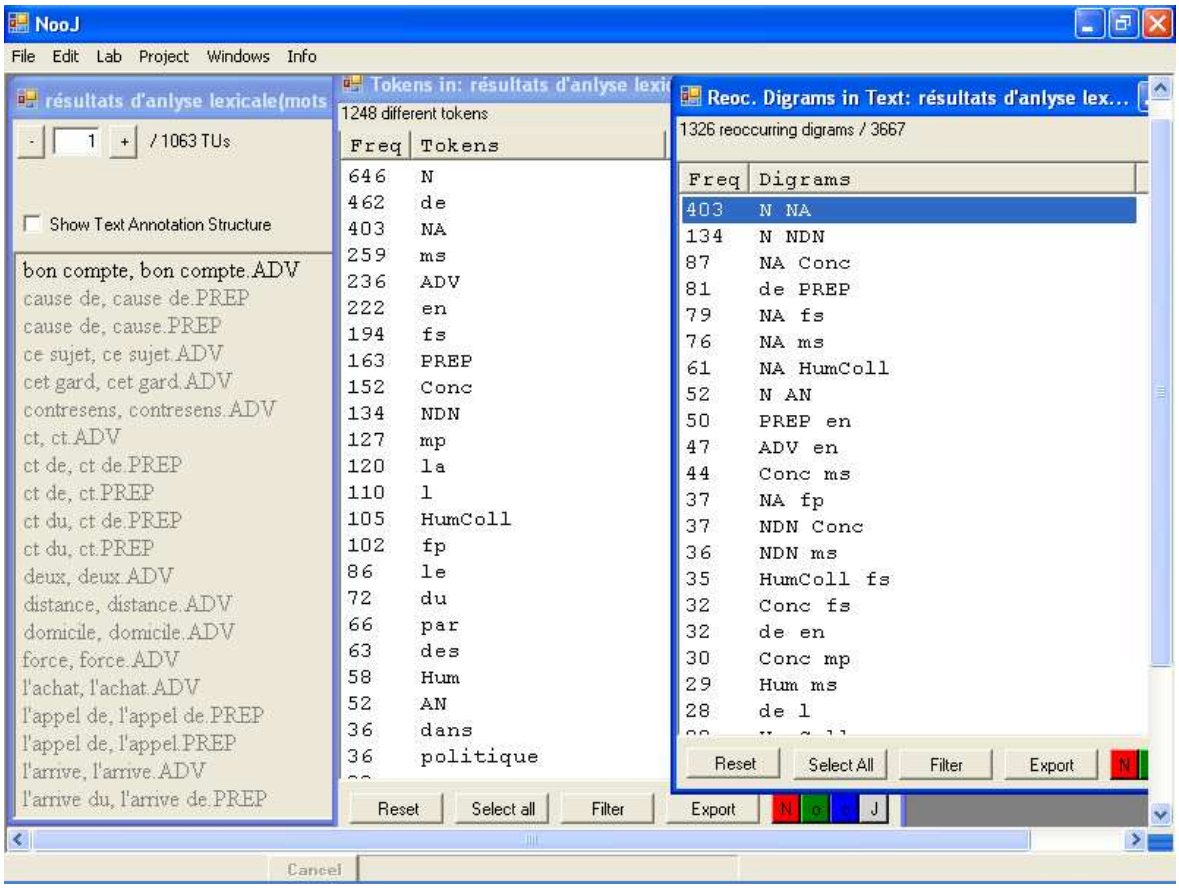


Figure.IV.19. *Résultat de l'analyse la liste des mots composé obtenue par Intex au moyen du logiciel Nooj*

Pour cette analyse nous avons utilisé deux fonctionnalités de cette application . La première est celles de *Tokens* qui est déjà utilisée pour la première liste , La seconde s'appelle *diagrams* ; elle nous permet de reconnaître les séquences de deux formes les plus fréquentes dans la liste . En observons la deuxième liste à gauche de l'écran ci-dessus , nous trouvons que les séquences les plus récurrentes sont respectivement *N NA* (403), N NDN(134), *NA Cons(87), NA fs(79), NA ms (76), HumColl (59), N AN (52), NA fp(37) etc*

Nous remarquons que les séquences le plus récurrentes sont les symboles de la déscription morphosyntaxique des entrées lexicales du dictionnaire DELACF obtenues au moyen de l'opération de l'étiquetage effectuée par le transducteur de ce dictionnaire (le programme automatique qui le régit). Alors, pour les comprendre, nous devons recourir au code de ce dictionnaire.

Les tableaux suivant nous expliquent le code d'étiquetage des dictionnaires DELAF :

Code	Signification	Exemples
A	adjectif	fabuleux
ADV	adverbe	réellement, à la longue
CONJC	conjonction de coordination	mais
CONJS	conjonction de subordination	puisque, à moins que
DET	déterminant	ses, trente-six
INTJ	interjection	adieu, mille millions de mille sabords
N	nom	prairie, vie sociale
PREP	préposition	sans, à la lumière de
PRO	pronom	tu, elle-même
V	verbe	continuer, copier-coller

1-Codes grammaticaux

Code	Signification	Exemple
z1	langage courant	blague
z2	langage spécialisé	sépulcre
z3	langage très spécialisé	houer
Abst	abstrait	bon goût
Anl	animal	cheval de race
AnlColl	animal collectif	troupeau
Conc	concret	abbaye
ConcColl	concret collectif	décombres
Hum	humain	diplomate
HumColl	humain collectif	vieille garde
t	verbe transitif	foudroyer
i	verbe intransitif	fraterniser
en	particule pré-verbale (PPV) obligatoire	en imposer
se	verbe pronominal	se marier
ne	verbe à négation obligatoire	ne pas cesser de

2- *Codes sémantiques*

Code	Signification
m	masculin
f	féminin
n	neutre
s	singulier
p	pluriel
1, 2, 3	$1^{ère}$, $2^{ème}$, $3^{ème}$ personne
P	présent de l'indicatif
I	imparfait de l'indicatif
S	présent du subjonctif
T	imparfait du subjonctif
Y	présent de l'impératif
C	présent du conditionnel
J	passé simple
W	infinitif
G	participe présent
K	participe passé
F	futur

3-*Codes flexionnels*

Tableau.IV.5. Tableaux des codes utilisés par les dictionnaires électroniques DELA[1]

Nous devons aussi, comprendre l'emploi des ces code la liste des mots composés , nous commençons par le code *N* qui signifie nom composé ; il est récurrent 646 fois 'd'après la liste des fréquence des mots *(tokens)* . Pour vérifier ceci nous relevons les occurrences de cette forme *N* au moyen de la fonction *concordance* de cette application . Nous sélectionnons la forme *N* puis nous cliquons sur l'onglet coloré NOOJ en bas, a droite de l'écran et voila le résultat obtenu :

[1] LEROI, Marie Véronique, Op.cit p. 128.

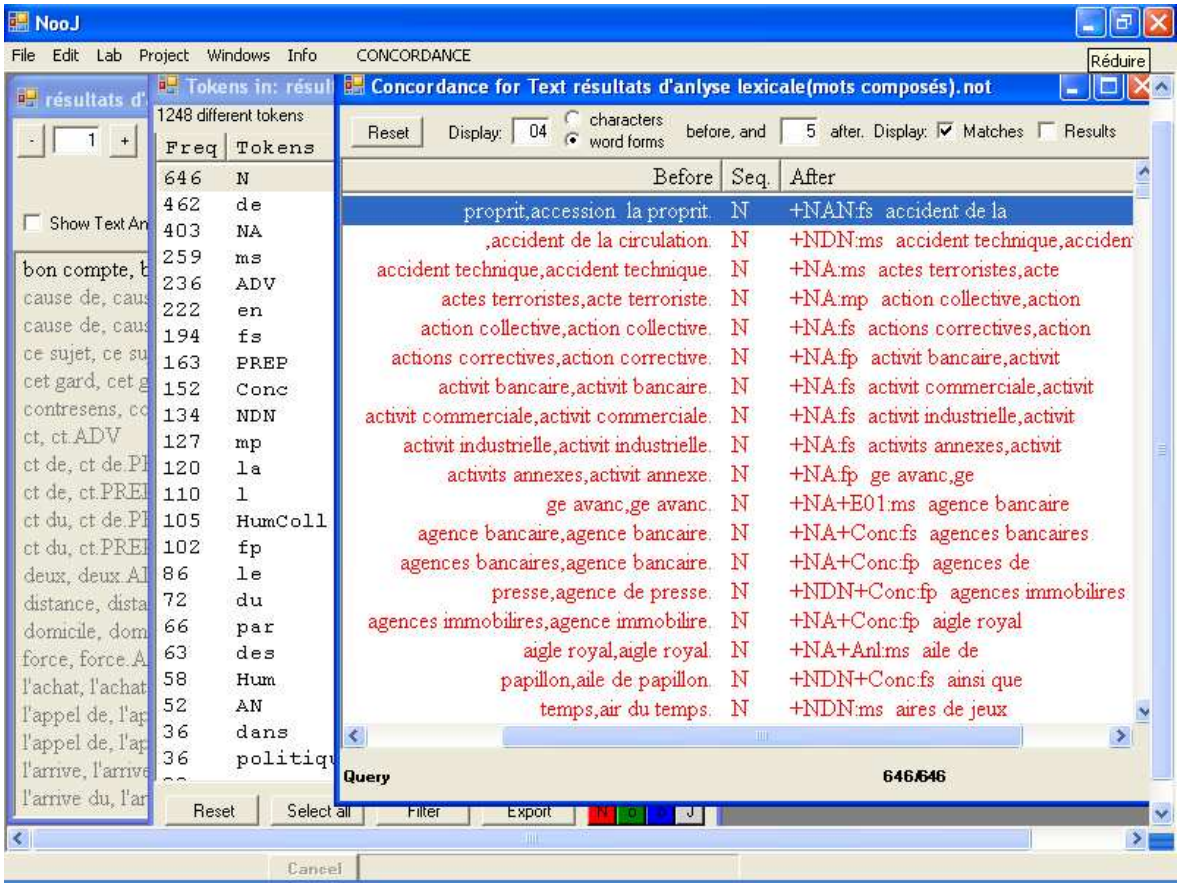


Figure.IV.20. *Les occurrences de la forme N dans la liste des mots composés*

3.3.5.3. Les types et les fréquence des expressions de la liste des mots composés (*compound words*) dans le corpus

Nous remarquons sur la listes ci-dessus, des occurrences de la forme *N* que le code N vient toujours après un nom composé puit nous avons un + suivi d' un autre code qui représente la constitution morphosyntaxique de l'expression. Exemple :

- *"accident de la circulation, accident de la circulation* N +NDNms ",

"N +NDNms" veut dire qu'il s'agit d'un nom composé constitué d' un nom, un déterminant et un autre nom.

- " *actions correctives, action corrective* N + NAfp" : nom composé dont a forme de base (le lemme) est action corrective ,composé d'un nom et adjective , féminin pluriel

En générale les résultats sont sous la forme suivante : *locution reconnue , le lemme co*rrespondant , *type de locution(N,ADV, PERP etc.) , constitution morphosyntaxique, flexion*

Ainsi, la fréquence des codes indiquant les types des expressions et nous permet de déduire la proportion de chaque type le corpus .En effet le résultat affiché sur l'écran ci-

dessus nous montre que le code N est répété 646 fois, ce qui veut dire que la liste des mots composé et bien entendu, le corpus contient *646 noms composés*. Ainsi nous déterminants le nombre de chaque types de mots composés reconnus dans cette liste

Pour bien confirmer la fiabilité de cette méthode nous essayons encore avec la forme *ADV* qui signifie une locution adverbiale et dont la fréquence est de 236 d'après l'écran ci-dessus .Pour ce faire nous relevons les occurrences de cette forme *ADV* en la sélectionnant dans la liste des mots fréquents et en cliquant sur le onglet rouge N en bas de l'écran . nous aurons, par la suite l'écran suivant :

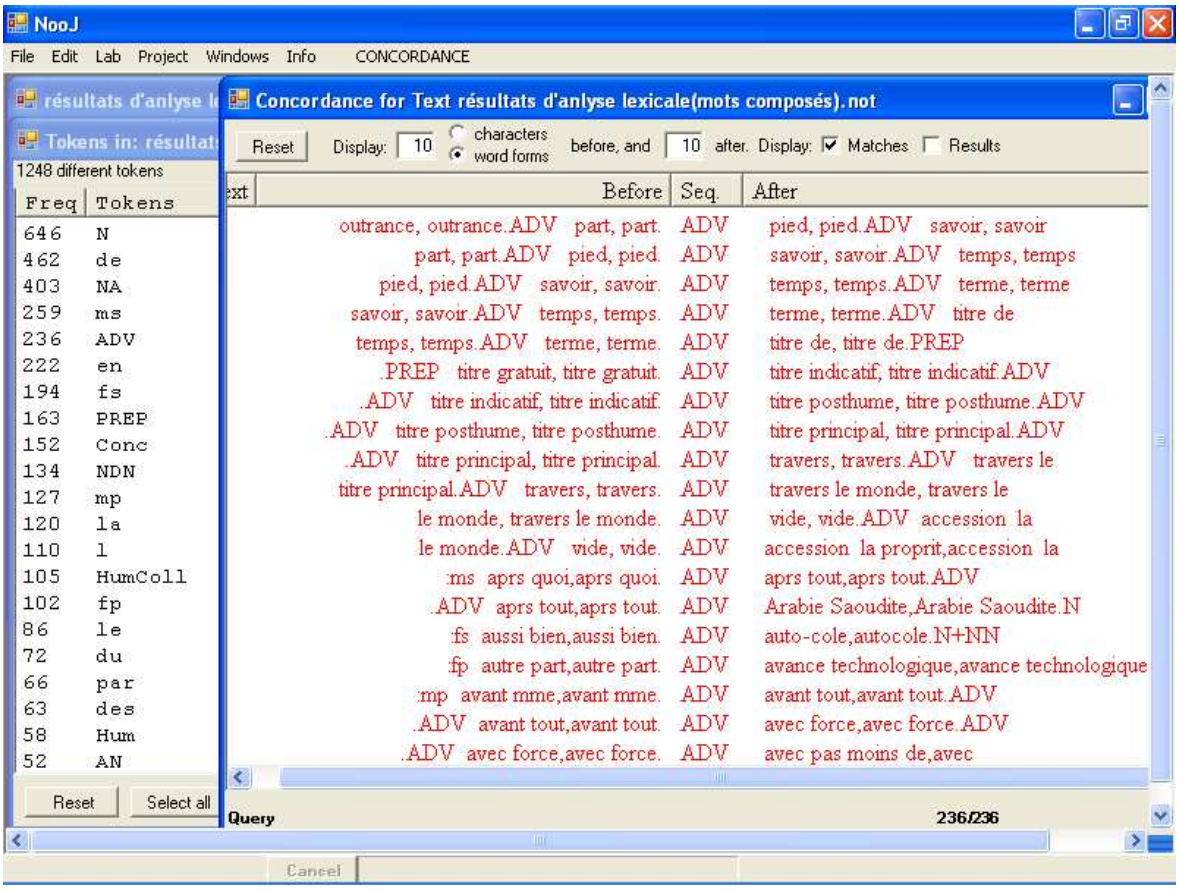


Figure.IV.21. *Les occurrences des locutions adverbiales dans la lise des mots composés*

Voila les occurrences des locutions adverbiales extraites au moyen de l'application Nooj mais nous remarquons au début de cette liste que les locutions *à part , à pied , à pied , à temps, à titre gratuit etc.* sont affichées sans la préposition à . Cela du à la non- reconnaissance du caractère *à* par l'application *Nooj* et il en va de même pour *é, è* qui font la différencie entre le code graphique du français et celui de l'anglais. Nous avons vérifié la configuration du programme; la langue française est choisie comme langue de

traitement mais le texte est les résultats du traitements figurent toujours de la même façon. Cependant ce défaut technique n' a pas constitué pas un obstacle pour la lecture et l'analyse des résultats,

Cette démarche nous permettra de faire la typologie d'une grande partie des expressions figées, employées dans notre corpus et déterminer leurs proportions, ce qui fait un des objectif de notre recherche. Nous pourrons, du coup éditer la liste de chaque type de locutions .

En premier lieux, au moyen de la même application Nooj, nous déterminons; dans la liste le nombre des occurrences de chaque type de mots composé :

Nous avons, déjà ci-dessus une partie la liste des **noms composés** et qui compte 646 expressions différentes et celle de 236 **locutions adverbiales** .Nous signalons ici que les occurrences de ces suites sont bien plus nombreuse parce chacune peut être employée plus d'une fois dans le corpus .

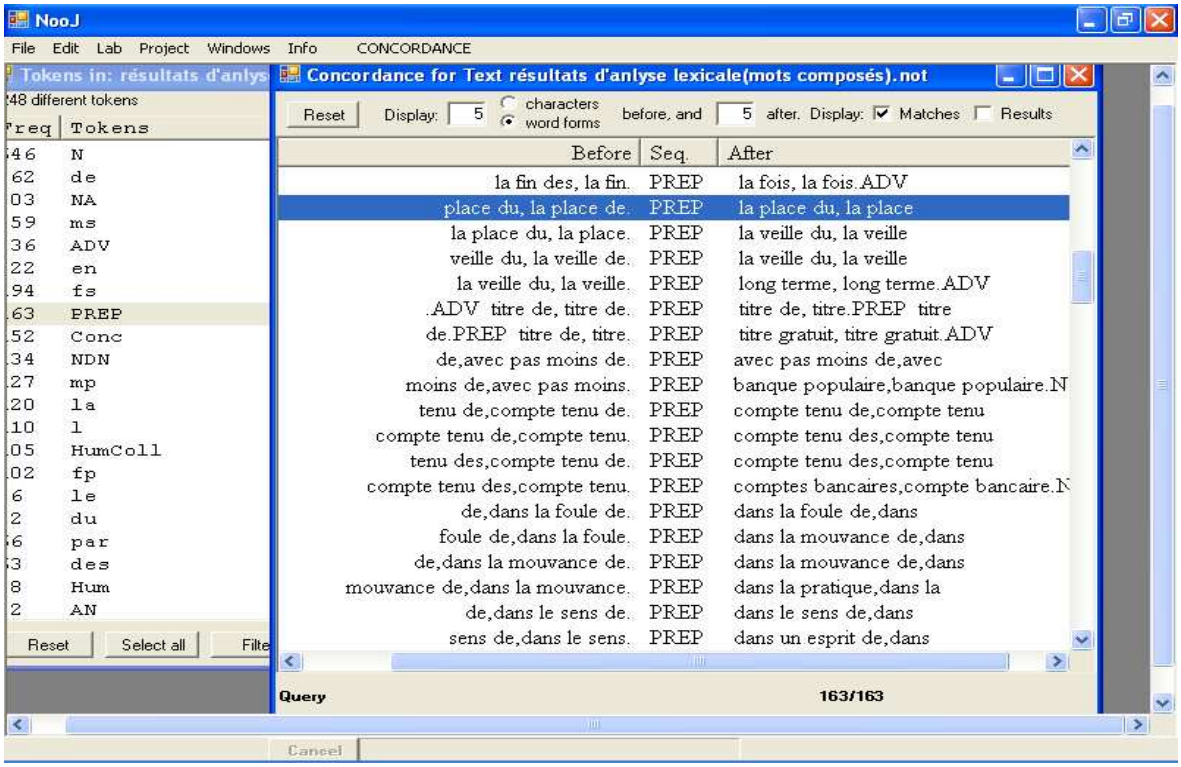


Figure.IV.22. les occurrences du code PREP dans la liste des mots composés.

Celle-ci est une partie la liste des occurrences du code PREP indiquant 163 **locutions prépositives** différentes employées dans le corpus et reconnues par l'analyse lexicale de Intex .Cette écran nous en affiche les suivantes : *à la fin de, à la place de, à la veille de, à titre de , avec pas moins de ,compte tenu de , dans la foule de , dans la mouvance de, dans le sens de* . Au début de la liste que la préposition *à* n'est pas

mentionnée à cause de la non reconnaissance de ce caractère par l'application *Nooj, ce* qui est du au problème déjà évoqué pour les locutions adverbiales.

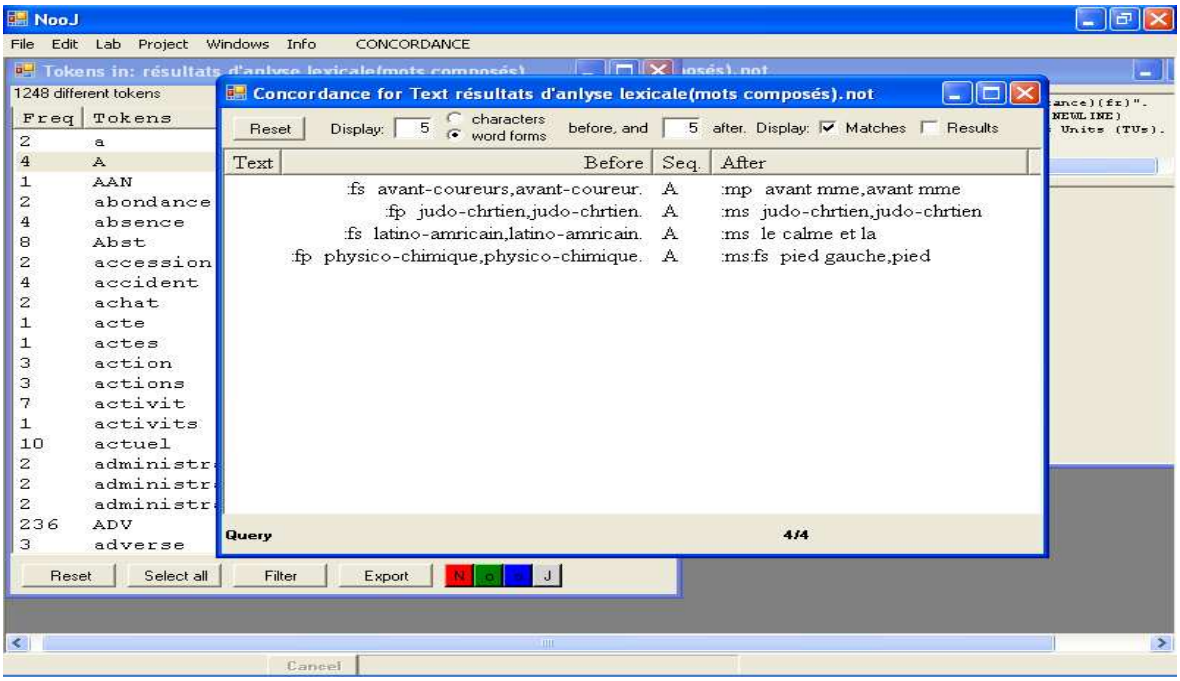


Figure.IV.23. Les occurrences des locutions adjectivales dans la listes des mots composés

Celle-là est la liste des **locutions adjectivales** qui contient 4 séquences indiquées par le code *A*

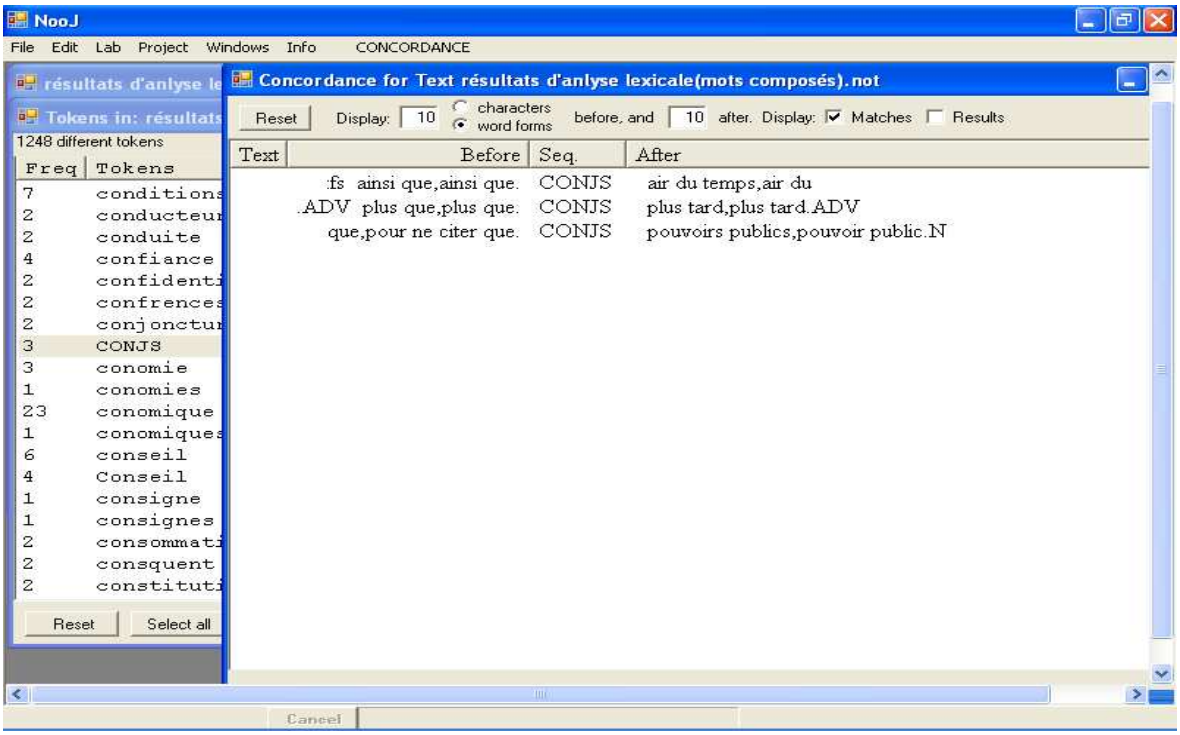


Figure.IV.24. Les locutions conjonctives reconnues dans le corpus.

Cet écran affiche les locutions conjonctives reconnues dans le corpus .

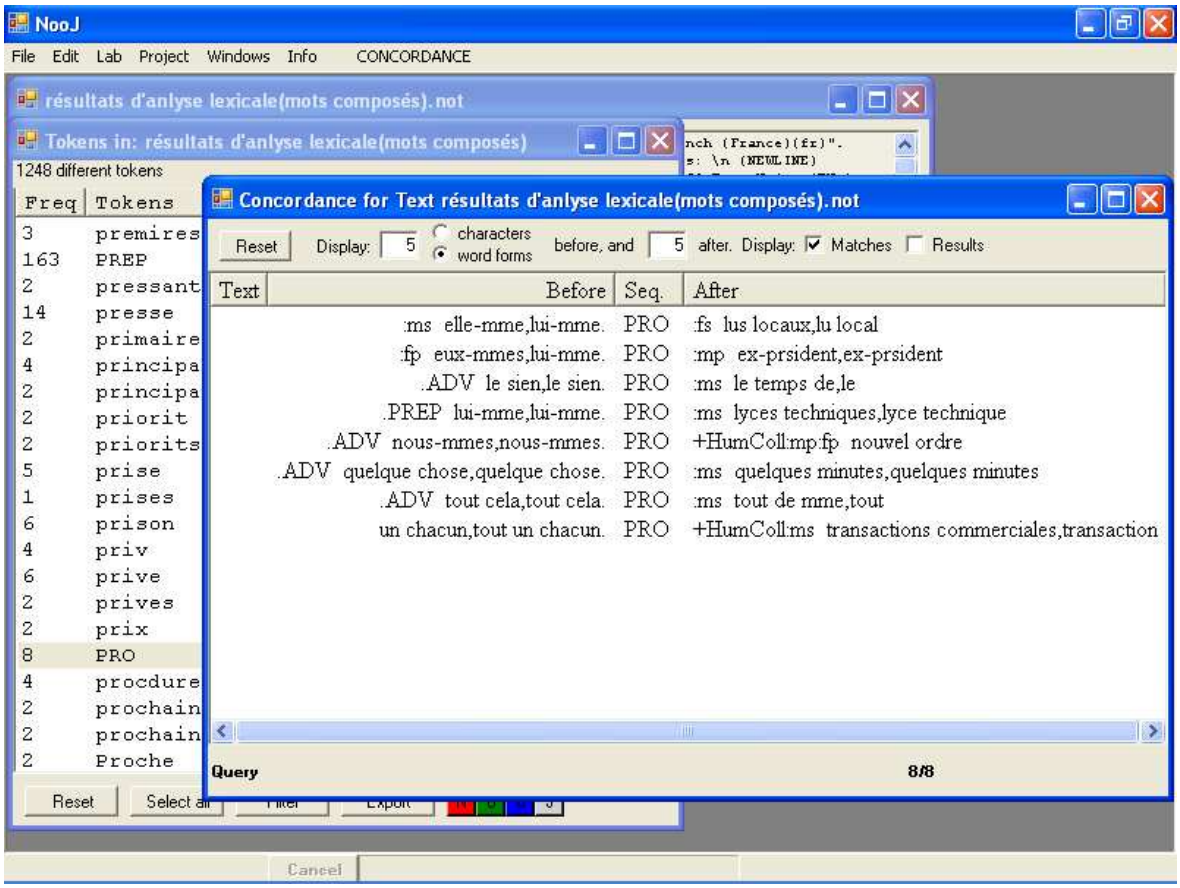


Figure.IV.25. Les pronoms composés contenus dans la liste des mots composés.

Quant à cet écran, il contient les différents pronoms composés employées dans ces extraite des journaux et reconnus par le dictionnaire DELCF en tant que mots composés et, bien entendu des expressions figées .

Ces résultats nous montrent que expressions les plus fréquentes et les plus variées sont les noms composés mais pour déterminer leur proportion par apport aux autres types de mot composé, nous devons compter les occurrences de chaque expression et en faire la somme . Cette tâche nous la rapportons à une autre phase de la recherche et nous nous contentons à ce niveau de faire la typologie des ces expressions, de relever les différentes locutions de chaque type et de compter les fréquence de quelques une à titre d'illustration.

Pour effectuer la dernière opération qui consiste à extraire et compter les occurrences de des locutions nous recourons à l'application *Intex* notamment la fonction *locate pattern* dans le menu *texte* . Nous cliquons sur l'onglet *texte* , le menu s'affiche ,nous choisissons *locate pattern* et nous aurons l'écran suivant :

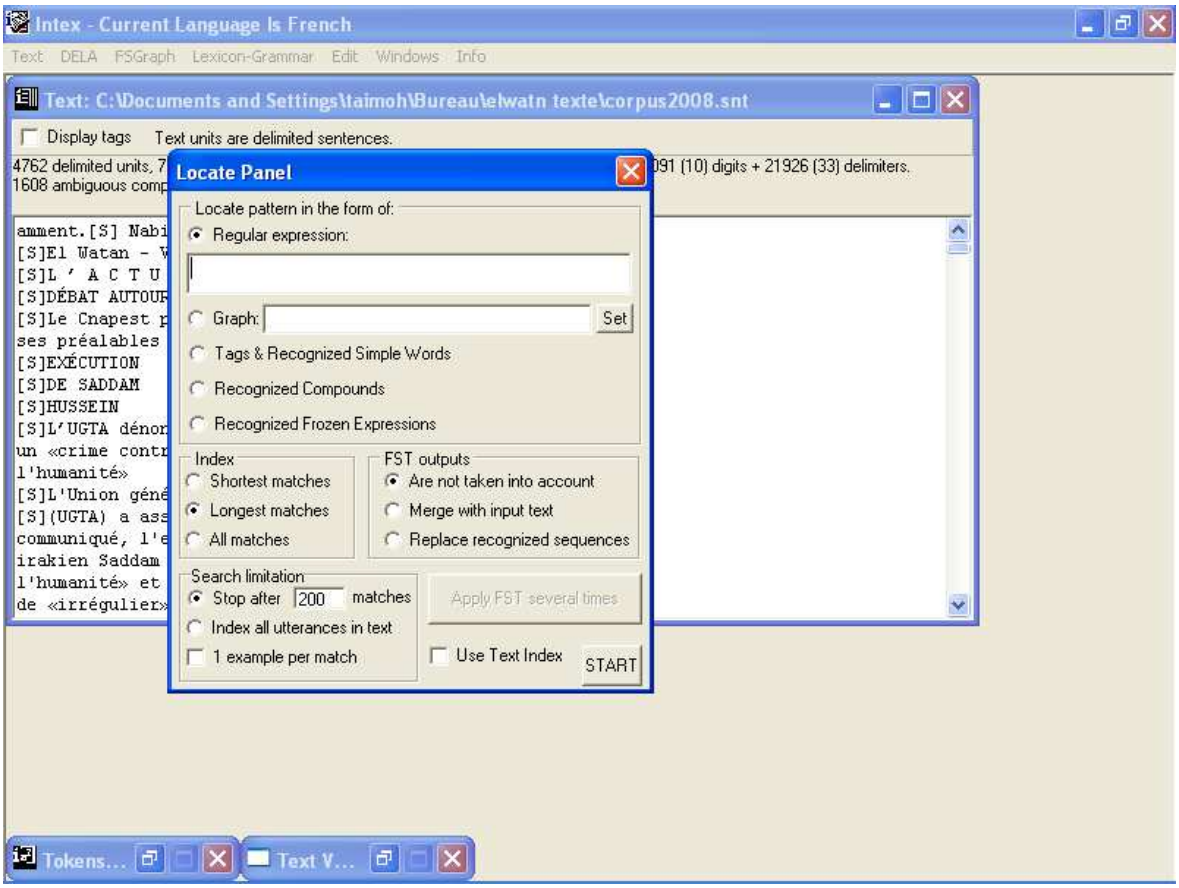


Figure.IV.26. Recherche des occurrences des expressions dans le corpus au moyen des logiciel Intex.

Nous prenons à titre d'exemple le nom composé *acte terroriste ;*. nous sélectionnons l'option *Regular expresions* ; nous tapons l'expression dans la barre d'écriture puis nous cliquons sur *start* .Voila le résultats que nous obtiendrons pour l'expression *agence de presse* :

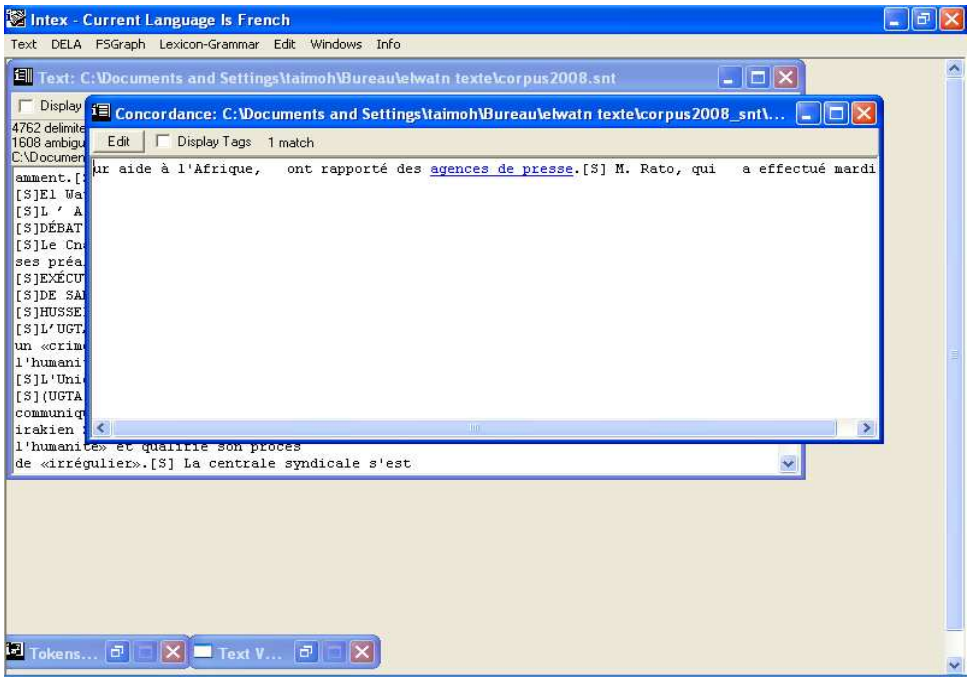


Figure.IV.27. Occurrences du nom composés *agence de presse* .

En continuant, ainsi cette quête nous trouvons que la plus part de noms composés ont une seule occurrence alors que les autres locutions sont bien plus récurrentes Par exemple la locution conjonctive "*ainsi que*" est employée, d'après Intex dans 23 occurrence et voila le résultat qu'elle donne :

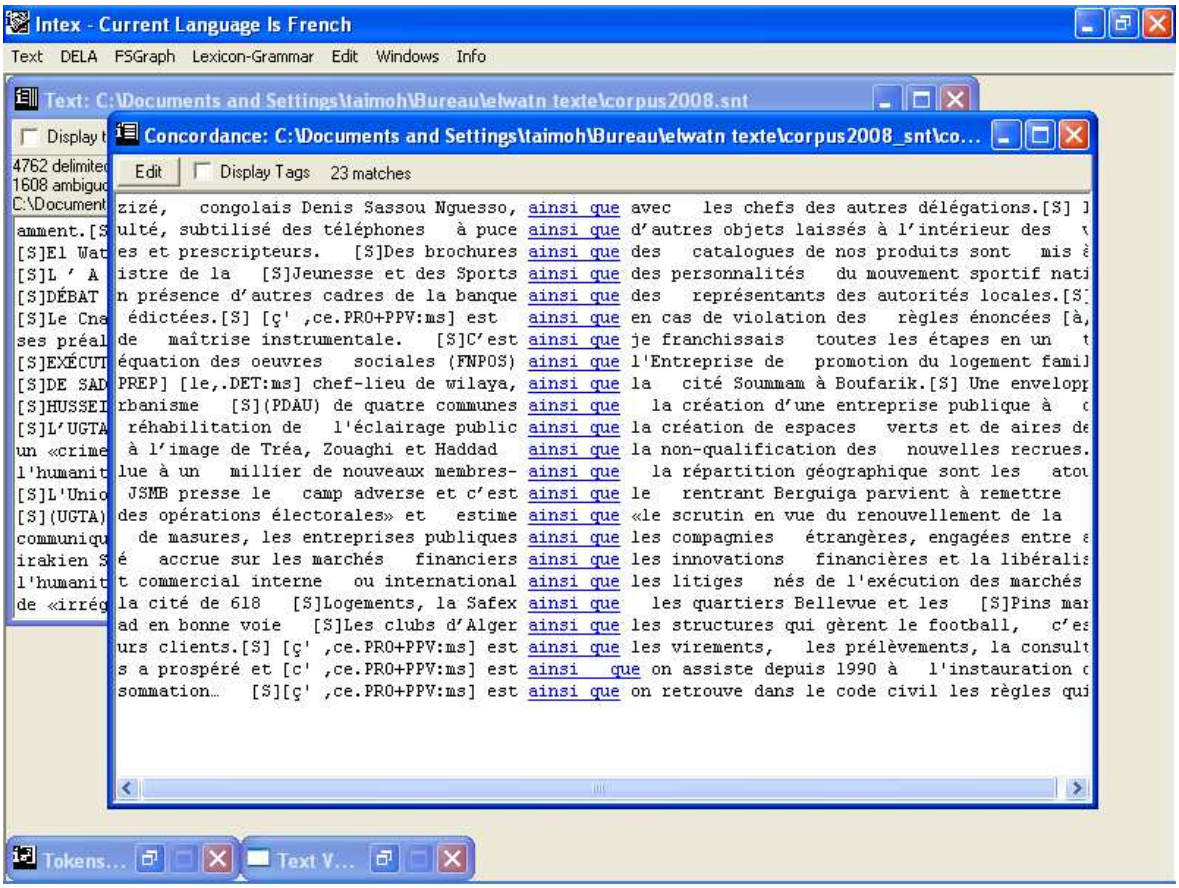


Figure.IV.28. les occurrences du mot composés *ainsi que*

Cette disparité entre les noms composé et les autres locutions de la fréquence d'mploi pourrait être du au caractère général des langues naturelles qui consiste à la supériorité des unité linguistiques grammaticales sue les unités lexicale . Cette propriété se révèle clairement dans la liste des mots (tokens) les plus fréquent dans le corpus .

Pour bien illustrer cette observation nous récapitulons les résultats de l'analyse des mots composés dans le tableau suivant :

Types des expressions	**Nombre des expressions différentes reconnues**	**Exemples des locutions**	**Fréquence de chaque expression dans le corpus.**
1- les noms composés	646	- *acte terroriste.* - *classe moyenne.* *-gamme de produits.*	01 03 02
2-les locutions adjectivales	4	- *avant-coureur.* *-judéo-chrétien.* - *latino-américain.*	01 02 01
2- les locutions	236	- *à cet égard.*	03

adverbiales		*- de surcroît.* *-par la suite.*	02 05
3- les locutions prépositives	163	*-à l'exception de.* *-en faveur de.* *- par le biais de.*	01 05 06
4- les locutions conjonctives	03	*-ainsi que.* *- plus que.* *- pour ne citer que.*	23 04 02
5- les pronoms composés	08	*-Lui-même.* *-tout cela.* *-quelque chose.*	07 01 04

Tableau .IV.6. Les différents types des mots composés reconnus dans le corpus.

Au cours, des recherches et des essais, nous avons trouvé que l'application Intex contient aussi, une fonctionnalité qui permet de compter et afficher toutes les occurrences de tout les expressions repérée par l'application des ressources lexicales correspondant aux mots composés , notamment le dictionnaire Delacf conçu par Blandine Courtois, en 2002. Lequel contient 248 885 entées de mots composés qui comportent 236 969 noms, 4 4732 adverbes, 4 356 prépositions, 2119 adjectives, 615(noun phrases), 50 pronoms, 44 conjonctions.

Pour accéder a cette fonction, nous activons l'onglet *Locate Pattern* dans le menu *Text*. Nous choisissons l'option de recherche *Recongnized Compounds* qui nous permettra de afficher toutes les occurrences des mots composé reconnus par l'analyse lexicale , déjà effectuée .

Voici, l'écran du choix d'option de recherche;

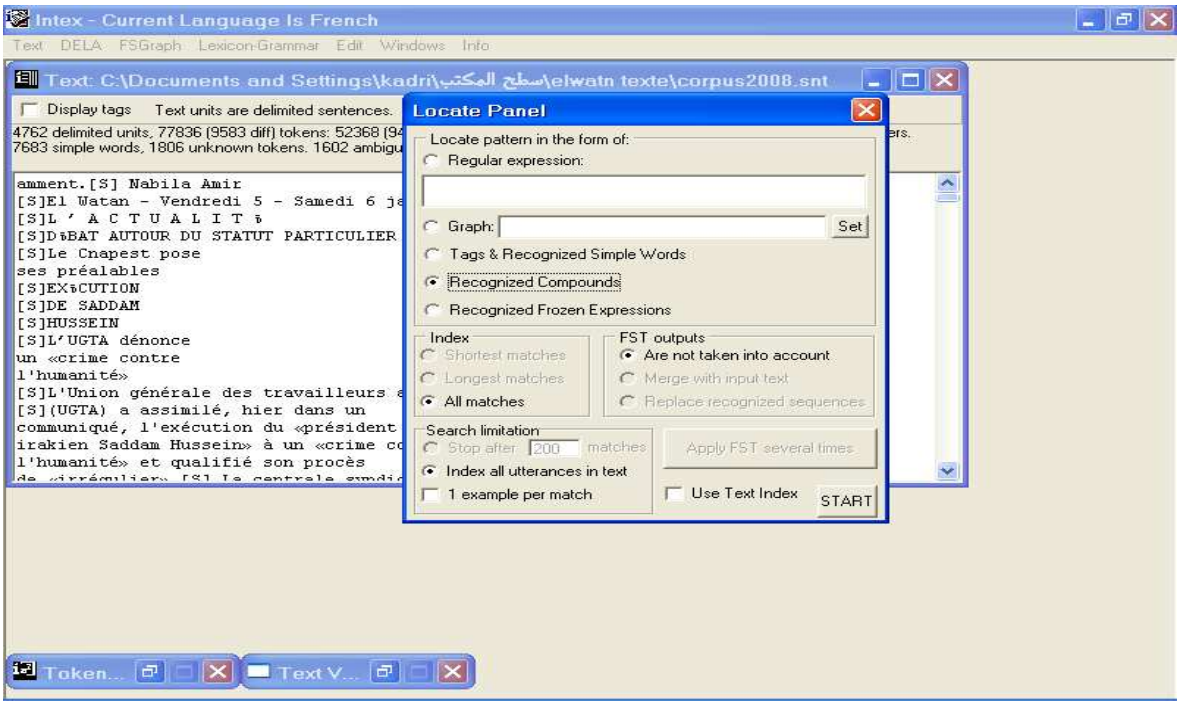


Figure.IV.29. Recherches des occurrences des mots composés reconnus par l'analyse lexicale.

Après que, nous choisissons l'option *Recongnised Rompounds* , nous lançons l'opération en appuyant sur l'onglet *STAT* , puis nous choisissons le nombre de mots que nous voulons avant et après chaque expressions , ensuite nous cliquons sur *Build concordance.* Après quelques secondes le résultat s'affiche dans l'écran suivant :

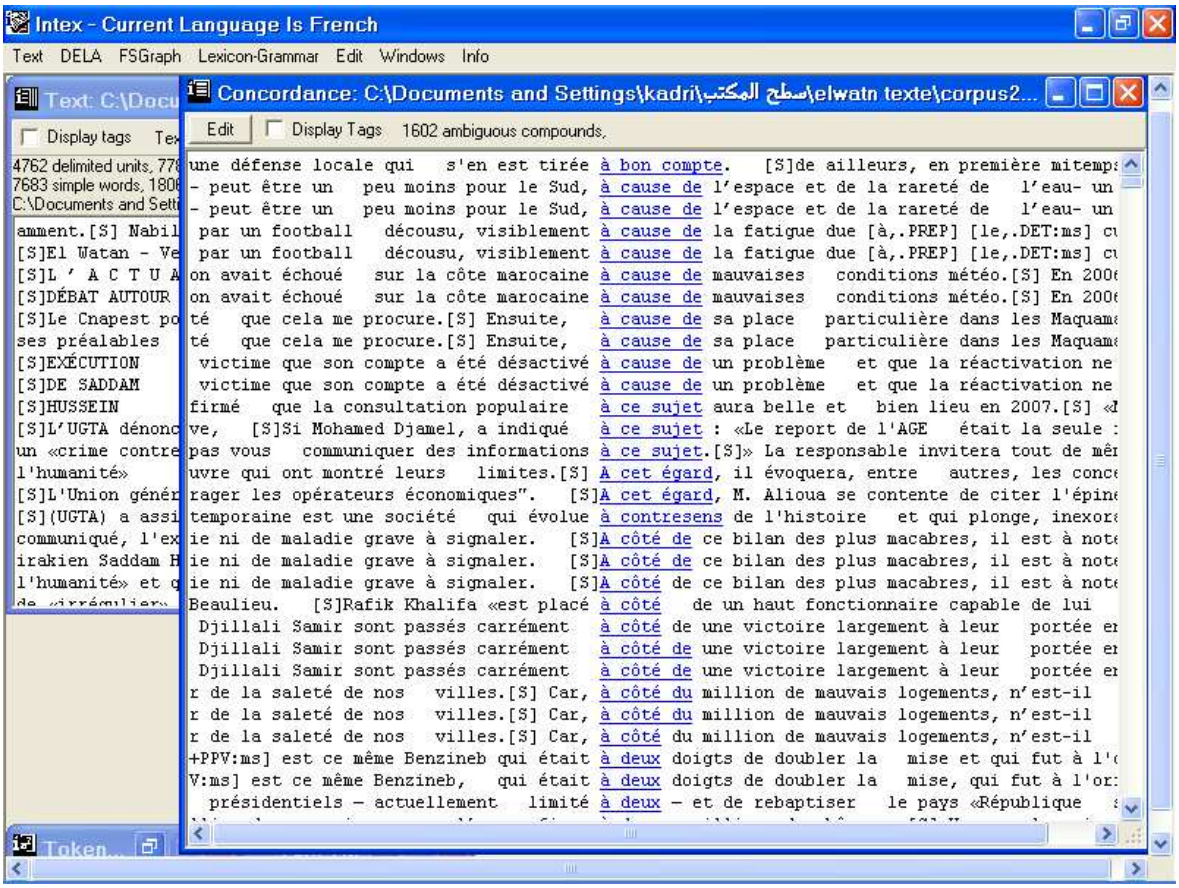


Figure.IV.30. résultat des occurrences de tous les mots composés reconnus dans le corpus.

Cet écrans nous montre que le système a compter 1602 occurrences de mot composés (c'est motionné au-dessus de l'écran), et nous affiche chaque fois le co-texte de l'expression employée dans le corpus. Alors que l'analyse lexicale a reconnu et compter les expressions différentes , c'est pourquoi , le nombre de cet résultat est relativement plus important .

3.3.5.4. L'importance des expressions figée dans les textes journalistiques

Nous supposons depuis le début de notre recherche que les expressions figées pourraient représenter un trait caractéristique pour les textes journalistiques. Affin de vérifier cette hypothèse nous recourons, également à notre méthode statistique automatique, en se servant de l'application Intext . Cette fois notre approche sera comparative; nous allons voir le nombre des expressions dans un corpus d'un autre genre des textes et nous en comparons la variétés et le nombre à la situations pour notre corpus de textes journalistique .

Nous avons choisi un corpus disponible sur l'application même. Il s'agit d'un roman qui s'intitule *la femme de trente ans* et voila un extrait de son texte :

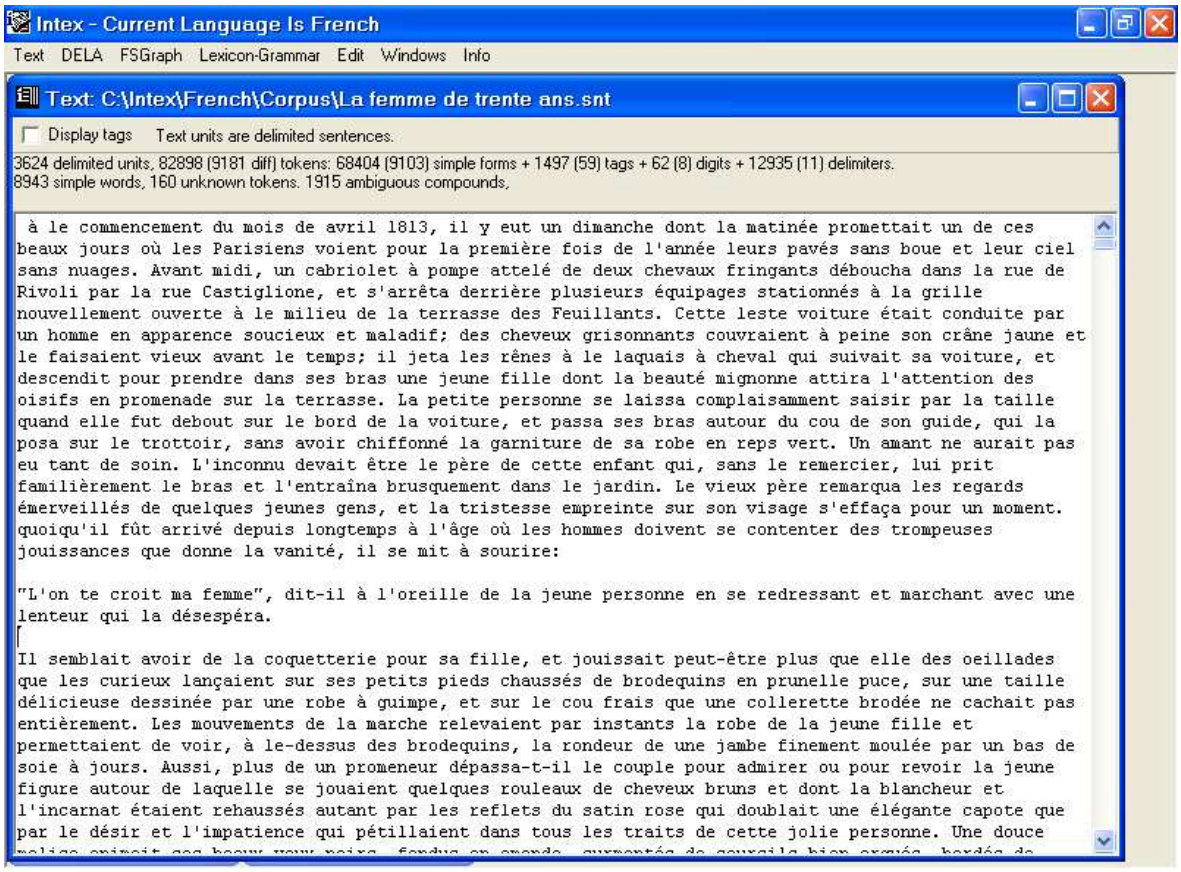


Figure.IV.31. extrait du corpus la femme de trente ans disponible sur l'application *Intex.*

L'écran ci-dessus nous affiche le texte et les informations initiales de ce corpus qui est composé, selon le prétraitement de Intex de 68404 formes simples dont 9103 différents . Nous effectuons, par la suite l'analyse lexicale de ce corpus au moyen de la fonction *Apply Lexical Ressources* et nous obtenons le résultat suivant:

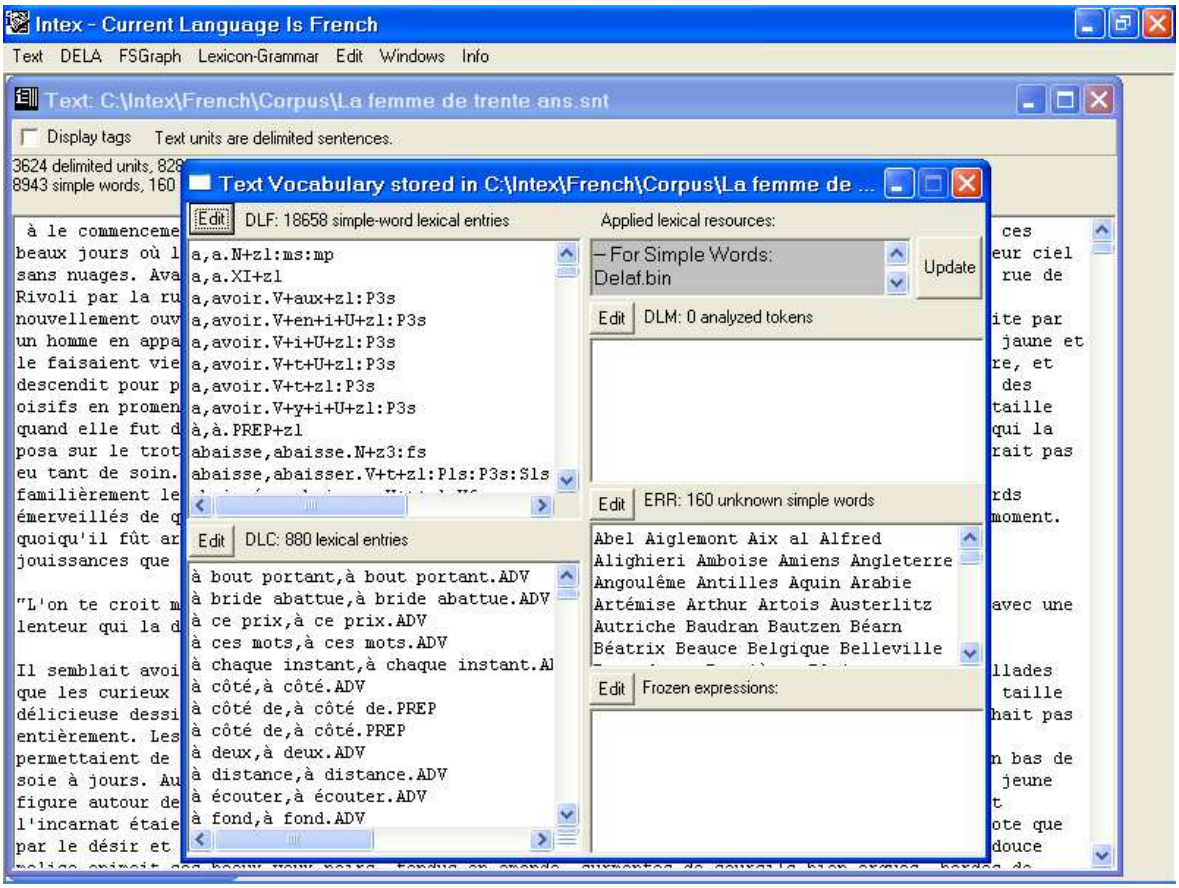


Figure.IV.32. Analyse lexicales du corpus la femme de trente ans

Cet écran nous montre que ce corpus contient 880 mots composés. Nous comparons cet résultat à celui enregistré pour le corpus des journaux à travers le tableau suivant:

	Les mots simples	Les expressions reconnues	Le taux
Le roman *La femme de trente ans.*	60404 mots	880 mots composés	68,64mots par mot composé
Les textes journalistiques	52368 mots	1062 mots composé	49,31 mots par mot composé.

Tableau .IV.7. Taux des mots composés par rapport aux mots simples dans les deux corpus

Ce tableau nous indique que les textes journalistiques sont plus riches en mots composés(noms composés, locutions adjectivales , adverbiales prépositionnelles et conjonctives et pronoms composés) que le texte du roman . En effet pour le texte du roman le système a reconnu un mot composé pour chaque 68,64 mots simples alors que dans les

textes journalistiques le présence des ces séquences représente une mot composé pour chaque 49,31 mots simples.

Nous concéderons ce résultat comme indicateur de l'avantage des expressions figées dans les textes journalistique par rapport aux d'autres genre de textes tandis que cette hypothèse n'est pas complètement confirmée .En effet la confirmation de telle hypothèse exige, nous semble-t-il plus de recherches empiriques portant sur toutes les formes des expressions figées et sur d'autres genre de textes .

4. conclusion

Au terme de ce chapitre ,nous soulignons, en guise de conclusion que de nombreux caractéristiques des expressions figées favorisent l'utilisation des méthodes du TAL comme outil de recherche dans ce domaine et dans le domaine de la phraséologie en générale . En effet ce phénomène correspond à des propriétés des suites lexicales, notamment la composition et la récurrence due à la mémorisation et la conventionalité des ces séquences.

De ce fait, nous avons voulu, à travers ce chapitre expliciter cette corrélation entre le figement et le TAL en expliquant les principaux outils et notions de cette méthodologie et en appliquant ces outils sur un corpus textuel .Notre approche s'est appuyé sur l'exploitations automatique des ressources lexicale évoluées , lesquelles ne consistent pas seulement à des listes des entrées des mots simples et leurs définitions mais il s'agit des systèmes opérant à partir des bases de données contenant en plus des différentes formes lexicales toutes les constructions phraséologiques reconnues par les connaisseurs et les locuteurs d'une langue donnée. Ces ressources assurent également des descriptions morphosyntaxiques et prennent en compte toutes les variations flexionnelles et transformationnelles que pourraient subir ces constructions. Cette étude était, en fait une démonstration de l'utilité des ces outils pour décrire la présence des expressions figées dans un corpus de textes journalistiques. Par ailleurs, nous nous sommes, aussi servis de ces outils pour souligner la spécificité de leur présence dans les textes journalistiques par rapport à d'autres genres de textes notamment les textes romanesques.

Conclusion générale

Nous avons essayé, à travers cette étude de répondre à la question : comment peut-on décrire l'emploi et les caractéristiques des expressions figées dans les textes journalistiques au moyen des outils du traitement automatique de la langue (TAL)? Nous avons cherché, également à confirmer l'hypothèse que ces expressions sont d'une présence privilégiée dans les textes journalistiques par rapport aux d'autres genre de textes .

Pour ce faire nous avons exploré, en premier lieu le champ conceptuel de ce phénomène dans les différentes disciplines de la langue affin de cerner la notion et les critères du figement. A partir de cette étude nous pouvons conceptualiser la notion des expressions figées en vertu de trois éléments principaux:

- L'opacité sémantique : Cette restriction s'oppose à la propriété compositionnelle qui représente l'un des fondements de la grammaire traditionnel. En effet, le sens d'une séquence n'est plus le produit des sens de leurs unités composantes. De ce fait les expressions figées constituent, dans la langue des unités fonctionnelles non-compositionnelles ou opaques.

- la restriction morphosyntaxique qui porte sur plusieurs propriétés transformationnelles dont jouissent les séquences libres telles que la passivation, la pronominalisation, l'extraction, la relativation, etc. Cela touche, également à certaines propriétés morphologiques et flexionnelles.

- La mémorisation : Nous nous sommes rendu compte à travers cette étude d'une caractéristique importante des expressions figées. C'est une conception du phénomène du figement qui montre qu'un syntagme peut acquérir le statut d'un titre ou d'une phrase historique ou rituelle lorsqu'il connaît un taux de répétition et de notoriété qui le transforme en une inscription mémorielle. Il s'agit également d'un critère de figement qui a fait l'objet d'étude de nombreux psycholinguistes comme Grunig, Swinney et Cutler, Gibbs Dwight Bolingre etc. Ces études ont démontré que le figement est l'effet de la représentation de certains syntagmes de langue dans la mémoire comme des unités lexicales au même rang que les mots et ils proposent plusieurs modèles pour cette représentation et pour le mécanisme de la reconnaissance et la production de ces séquences.

Dans un second temps, nous avons montré, à travers cette étude, la pertinence de la méthodologie automatique pour la description des expressions figées dans la langue en générale et dans le discours journalistique en particulier. Nous nous sommes appuyés sur les recherches du LADL(le Laboratoire d'Automatique Documentaire et Linguistique),

crée en 1968 par Maurice Gross. Ce linguiste et ingénieur de formation et lexicologue a entrepris avec son équipe un vaste programme de description systématique des propriétés syntaxiques de tous les éléments du lexique français. Suite à son long séjour aux Etats-Unis, il a été influencé par les théories de Noam Chomsky mais il s'est appuyé, beaucoup plus sur la théorie transformationnelle de Z Harris qui parle des phrases élémentaires ou noyaux qui sont invariables et pourraient être considérées comme unité sémantique de base et non pas le mot. La nécessité de décrire de telles constructions avait conduit Z. Harris et Maurice Gross à adopter La théorie du Lexique-grammaire ou lexique syntaxique .

Dans le cadre de cette théorie et dans l'optique de décrire ces constructions élémentaires, Le LADL a élaborés des tables, appelées les tables de lexique-grammaire ; il s'agit des matrices pour représenter le comportement distributionnel et transformationnel des prédicats dans les phrase simples. Les lignes représentent les entrées lexicales comportant, en plus des mots simples des locutions et des expressions. Quant aux colonnes, elles représentent, chacune une propriété morphosyntaxique. Ces tables sont, par la suite systématisés et automatisés sous forme des dictionnaires électroniques permettant de reconnaître et décrire, dans un corpus donné les unités simples ou composés prédéfinies dans ces tables.

Nous avons essayé, dans notre étude de se servir des outils du TAL et notamment des réalisations de la théorie de lexique-grammaire pour étudier les expressions figées dans les textes journalistiques. En effet nous avons procédé à une méthode qui s'appuie sur l'application automatique des ressources lexicales électroniques. Ces ressources lexicales sont constituées à base des tables de lexique-grammaire décrivant les locutions et les expressions figées en l'occurrence la table C1d des locutions verbales figés. Pour ce faire nous avons opté pour un logiciel approprié à ce genre de traitement; il s'agit de l'application *Intex* et sa version évoluée *Nooj*. Cette recherche nous a permis, aussi de bien comprendre les principes et les fonctionnalités des ces applications informatiques et de montrer, par conséquent l'intérêt des outils du TAL pour l'étude du phénomène des expressions figées.

Le logiciel *Intex* et les ressources lexicales disponibles sur cette application sont conçus par Max Silberztien dans le cadre des recherches du LADL. Nous avons exploité pour notre recherche plusieurs fonctionnalités de cet outil informatique notamment la fonction de l'analyse lexicale que l'on pourrait activer en cliquant sur l'onglet *Apply lexical ressources* dans le menu *Text*. Cette fonction nous a permis d'appliquer, sur notre corpus,

le dictionnaire électronique des mots composés Delacf contenant de différents types de locutions et de noms composés telle que la table de lexique-grammaire C1d des locutions verbales laquelle est muni des programmes secondaires appelés *transducteurs* pour reconnaître toutes les variations et les flexions que pourraient subir ces expressions.

Les résultats de cette opération étaient, en premier temps sous forme des liste des séquences suivies chacune d'un code décrivant son type et éventuellement sa constitution morphosyntaxiques. Nous nous sommes servi de ce code et d'une autre application dénommée *Nooj* et développée à partir de *Intex* pour déterminer la fréquence de chaque type d'expression dans le corpus, notamment dans la deuxième liste des noms composées et des locutions. Cette analyse nous a permis d'établir la typologie des ces séquences dans le corpus et de rendre compte de la proportion de chaque type dans le corpus.

Par ailleurs, nous avons procédé à une deuxième analyse des cette liste qui nous a permis de compter et afficher toutes les occurrences des différentes expressions repérer par l'application des ressources lexicales. Nous avons utilisé, cette fois une fonction sur le logiciel Intex, appelée *Locat pattern* qui sert à rechercher et afficher les contextes, dans le corpus où ils s'emploient ces expressions (occurrences). Nous avons trouvé que notre corpus, constitué de 52368 mots simples contient 1062 locutions et noms composés, ce qui revient à un mot composé pour chaque 49,31 mots simples.

Nous avons constaté, ensuite que ce dernier résultat représente le taux de la présence des ces expressions dans notre corpus de texte journalistique, ce qui nous a donné l'idée de comparer cette situation à un corpus d'un autre genre des textes. Nous avons choisi pour ce faire un corpus de texte littéraire, en l'occurrence un roman disponible sur le logicien *Intex* qui s'intitule *La femme de trente ans*. Le résultat pour ce corpus était: un mot composé (expressions) pour chaque 68.64 mots simples, ce qui fait un taux de (1/68.64), lequel est beaucoup plus moins que celui du corpus des textes journalistiques. Nous nous sommes rendu compte, enfin que les résultats de cette dernière approche pourraient confirmer notre hypothèse que les textes journalistiques sont plus riches en expressions figées que d'autres genres de textes.

Nous voudrions souligner, en outre que cette étude n'était point sans écueils, ainsi sur le plan conceptuel qu'au niveau des méthodes du traitement du corpus. En effet les concepts d'expressions figées et de figement étaient en concurrence terminologique avec de nombreux d'autres termes. De fait, ces deux vocables se confondent souvent à d'autres

concepts tels que *phraséologie, idiomatique, locution, mot composé,* lesquels recouvrent des domaines qui s'imbriquent et s'entremêlent.

De plus, la signification des ces séquences pose, aussi un sérieux problème, ce qui est dû au caractère arbitraire, et à la non-compositionalité de leur sens et la pénurie des dictionnaires spécialisés en expressions et locutions figées.

Il en va de même pour les méthodes et les outils du traitement. Tout d'abord l'assimilation des ces méthodes était un laborieux apprentissage à partir de l'abc de ces techniques. Ensuite l'acquisition des ces logiciels exigeait une recherche épineuse dans le nombre très limité des ouvrages et des ressources du web disponibles. En outre le maniement des applications informatique a exigé une recherche minutieuse dans les manuels. D'ailleurs nous étions contraints de recourir à la traduction avec tous les ennuis qu'elle suscite puisque la documentation à ce sujet et les interfaces des logiciels étaient en majorité en anglais. Ces difficultés ont eu un fort impact sur la performance de quelques fonctionnalités et la fiabilité de certains résultats.

Par ailleurs, nous soulignons, en guise de conclusion que cette étude laisse apparaître d'autres perspectives de recherche dans ce domaine, ainsi dans ses dimensions linguistiques que sur le plan du traitement automatique. Nous signalons en ce terme que l'on pourrait étudier le phénomène du figement dans une perspective générative or, les expressions figées ont, souvent servi d'argument contre la grammaire générative et transformationnelle. En effet nous pensons qu'il serait pertinent d'étudier les mécanismes et les compétences qui permettent de générer les énoncés constitués d'expressions figées et s'interroger sur leur position vis-à-vis les grammaires qui assurent la génération du discours d'après la théorie chomskyenne. Sur le plan des méthodes automatiques nous proposons également de mener des études qui visent à développer des programmes qui permettront la reconnaissance et l'analyse des suites de mots de plus de deux lexèmes (*n-gram*) les plus fréquents dans un corpus à l'instar de la fonction *diagram* dans l'application application *Nooj* qui permet seulement la détection des suites de deux lexèmes. Un tel outil pourrait détecter les séquences en cous de figement dans un discours donné.

A ce point, nous arrivons au terme de ce travail mais nous ne pourrions point prétendre aboutir au terme de l'étude et de l'analyse de la problématique proposée dans le cadre de cette recherche. En effet nous n'avons qu'à reconnaître que le figement, en tant que fait linguistique et phénomène discursif caractérisant toutes les langues naturelles est loin d'être cerné par une étude de dizaine de pages. De ce fait, ce travail ne pourrait guère

être loin de toutes insuffisances et lacunes, surtout lorsque nous abordons un sujet de telle consistance et nous procédons à une méthodologie récente qui exige une certaine maîtrise technique et scientifique de l'outil informatique.

Nous concluons en définitive que ce travail confirme, nous semble-t-il l'intérêt des méthodes automatiques pour l'étude du phénomène du figement, ainsi que l'intérêt de l'étude des expressions figées, comme étant une caractéristique inhérente aux langues naturelles, pour la description de la langue et du langage.

Références bibliographiques

Ouvrages

1- Amossy, R, Herschberg Pierrot, A., *Stéréotype et clichés*, Arman colin, France, 2005**.**

2- Anscombre, J-C, Proverbes et formes proverbiales: valeur évidentielle et argumentative,

Langue française, n°102, pp.95-107, 1994.

3- Anscombre, J-C, *Théorie des topo*i, Kimé, Paris, 1995.

4- Bailly, C., *Traité de stylistique française*, Librairie Georg, Paris, 1951.

5- Bernard G, Les locutions verbales françaises, *La linguistiqu*e, Paris 1974, pp. 5-17,

6- Catach, N, *Orthographe et lexicographie*, Nathan, 1981.

7- Charaudeau, P, Maingueneau, D, *Dictionnaire d'analyse du discours*,Le Seuil, Paris, 2002.

8- Courtois B, Silberztein M., Les dictionnaires électroniques du français, *Langue Française*, n° 87, Larousse, Paris, 1990

9- Danlos L, La morphosyntaxe des expressions figées , *Langages*, n°63. 1978

10- Dournon J. *Le dictionnaire des proverbes et dictons de France*, Hachette, Paris, 1986.

11- Duboi J. et al. *Dictionnaire de linguistique*, Larousse, Paris, 1973.

12- Ducro O, Todorov T., *Dictionnaire encyclopédique des sciences du langage* . Seuil, Paris, 1979.

13- Dugas A, La création lexicale et les dictionnaires électroniques. *Langue Française*, 87, pp. 23-329, 1990.

14- Fiala, Pierre et al , *La locution : entre lexique, syntaxe et pragmatique. Identification en corpus, traitement, apprentissage.* INALF, collection Saint-Cloud, Klincksieck, Paris,1996.

15- Gross, G, *Degré de figement des noms composés*, Langage, n°90, 1988

16- Groos G., *Les Expressions figées en français*, Ophrys, France, 1996.

17- Groos G., « *Du bon usage de la notion de locution* » , La locution entre langue et usages, ENS éditions, Fontenay Saint-Cloud.1997.

18- Gross M, *Méthodes en syntaxe*,Hermann, 1975.

19- Gross M., *Une classification des phrases figées du français*, In: *De la syntaxe à la pragmatique* , ATTAL P. et MULLER C.., John Benjamins publishing company, pp. 141-180, 1984 .

20- Gross, M, « Les limites de la phrase figée », *Langages*, Larousse, Paris. n°90, 1988.

21- Gross M., Quelques réflexions sur le domaine de la traduction automatique, *TAL*, Paris, 1992.

22-Hudson, J , *Perspectives on fixedness: applied and theoretical*, Lund Studies in English n°94, Lund University Press, Lund. 1998 .

23-Kleiber, G , la sémantique du prototype. Catégories et sens lexical, Presses universitaires de France, Paris. 1990 .

Lafleur, B., .*Dictionnaire des locutions idiomatiques françaises*, Duculot, Ottawa, 1979.

24- Leclere C., Organisation du Lexique-Grammaire des verbes français, Langue Française, n°87, Paris, Larousse. 1990.

25-Maingueneau D., *Analyse des textes de communication* , Arman Colin , Paris, 2005.

26-Marchand B, *L'analyse du discours assistée par ordinateur*, Armand Colin, Paris, 1998.

27 - Marouzeau. J., *Lexique de la terminologie linguistique*, Librairie orientaliste Paul Geuthner, Paris, 1962.

28- Martins-Baltar, M , *La locution entre langue et usages*, ENS éditions, Fontenay Saint-Cloud, 1996

29- Mejri S., Séquences figées et expression d'intensité. Essai de description sémantique, *Cahiers de lexicologie*,n° 65, pp.111-122, 1994.

30- Misri, G, Approches du figement linguistique : critères et tendances, *La linguistique* , Vol 23 ,Paris, pp. 71-85, 1987.

31-Moon R, *Fixed expressions and idioms in English, a corpus-based approach*, Clarendon press, Oxford. 1998.

32- Mortureux M., *la lexicologie entre longue et discours*, Arman Colin , Paris, 2004

33- Rey A, Chantreau S, *Dictionnaire d'Expressions et Locutions*, Dictionnaires Le Robert, Paris, 1989.

34- Salkoff M., *Une grammaire en chaîne du français : Analyse distributionnelle*, Paris, Dunod, 1973.

35- Schapira, C , *Les stéréotypes en français : proverbes et autres formules*,Ophrys. 1999.

36-Senellart J., Reconnaissance automatique des entrées du lexique-grammaire des expressions figées, In : *Le lexique-grammaire*, Lamiroy B., Travaux de linguistique, Bruxelles, 1999.

37- Silberztein M., *Dictionnaires électroniques et analyse automatique de textes : Le système INTEX*, Masson, Paris, 1993.

38- Silberztein M., Transducteurs pour le traitement automatique des textes, *Le lexique-grammaire*, Lamiroy B. , Travaux de linguistique, Bruxelles, 1999.

Sitographie

- Constant, Mathieu .*Vers la construction d'une bibliothèque en-ligne de grammaires linguistiques,* , Université de Marne-la-Vallée. http://www.ladl.univ-mlv.fr2002.
- LE ROI, Marie-Véronique, *Traitement automatique et lexicographique des locutions verbales figées en français*, mémoire soutenu à l'université Paris III, Sorbonne nouvelle ILPGA. http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/sitespp/maitrise-2004/slMVLeroi- 2004.pdf.
- SEVNSSON, Maria Helena , *Le critères de figement .L' indentifications des expressions figées en français contemporain, ,* Print &Media, Umeá, 2004. http://www.duo.uio.no/roman/Art/Rf-16-02-2/fra/Svensson.pdf.
- *Enjeux linguistiques et informatiques des expressions figées* . http://www.limsi.fr/Individu/habert/Publications/Fichiers/habert91b/BH_C1.html.
- CLAIRE,Gardent, BRUNO, Guillaume,lexique syntaxique et tables du LADL, CNRC/LORIA, Nancy, France 2006. http://www.exsynt.inria.fr/talk/parisSept05.pdf.
- CLAIRE Gardent BRUNO, Guillaume, *Extraction d'information de sous-catégorisation à partir des tables du LADL*, Nancy . http://www.loria.fr/~perrier/taln06.pdf .
- SILBERZTREIN, Max , *Manuel INTEX* . http://www. mshe.univ-fcomte.fr/intex/downloads/Manuel.pdf,
- Claudia Maria Xatara, *Les expressions idiomatiques : de la marginalité à la reconnaissance,* Université de l'Etat de Sao Paulo, Brésil . http://fdlm.org/fle/article/319/idiomatique.php3
- Sébastien Paumier, *Dictionnaire électronique des mots composés (DELAC),* http://infolingu.univmlv.fr/DonneesLinguistiques/Dictionnaires/delac.html

- Pierre Dupont, *Localisation et analyse d'expressions figées dans de grands corpus,* Cédrick Fairon (CENTAL) . http://www.info.ucl.ac.be/enseignement/memoires/2003-2004/pdupont_tal1.html

Corpus informatisé

Le journal quotidien El Watan: http:// www.elwatan.com/

Les numéros : 04-01-2007
06-01-2007
08-01-2007
09-01-2007
10-01-2007
11-01-2007
14-01-2007
13-01-2007
17-01-2007
20-01-2007
21-01-2007
. 22-01-2007
23-01-2007
24-01-2007
30-01-2007
31-01-2007

Supplément économie, N°88.

Supplément immobilier,N°44.

Supplément T-V N°25 .

Logiciels

-Le logiciel *Intex.*

http://intex.univ-fcomte.fr/downloads/

- L'application *Nooj*

http:// www.nooj4nlp.net/

Dictionnaires

Le Larousse expression [CD-ROM. (2002)] Paris : Larousse.

Liste des tableaux

Liste des schémas

Liste des figures