Problèmes soulevés par l'étiquetage des corpus littéraires

Katia ZELLAGUI et Jean-Marie VIPREY

Abstract : Our laboratory team (the Grelis Group Of Research in Linguistic, Computer Sciences et Semiotic) of the University of Franche-Comté works in computer treatment of literary corpora. In our collaboration with Intex community, we propose to bring out new tools for tagging literary texts. In the tagging process, after the segmentation of the text in sentences, we need in a first time to identify and tag the compound forms. As we know, a compound noun can be ambiguous. Intex can identified in the text all compounds included in the DELAC and in the DELACF... that means first that the compound non-included in these dictionaries won't be identified and tagged, and second that an identified sequence can be ambiguous. We propose in this article a way to identify the compounds and the recurrent sets specific to the author in a literary corpus. We'll filter through the obtained lists by using current dictionaries (the DELAC for example) and we'll analyse the non-identified sequences. If a sequence presents the characteristics of compound, we'll include it in a specific dictionary: a corpus dictionary.

Keywords: Intex, local gramar, electronic dictionaries, statistical analysis, hypertext.

Mots clés : Intex, grammaires locales, dictionnaires électroniques, analyses statistiques, hypertexte.

1. Introduction

Nous proposons ici une procédure d'aide à l'identification des formes composées dans un corpus. Nous bénéficions de l'accès à une version scientifiquement établie de *La Comédie humaine*, corpus de 4,2 millions d'occurrences « simples » (et 630000 ponctuations), 4,7 carac-

✉ Katia ZELLAGUI, Jean-Marie VIPREY, Laboratoire GRELIS,
30, Rue Mégevand, 25035 Besançon e-mail : katia.zellagui@univ-fcomte.fr
jmviprey@utinam.univ-fcomte.fr

tères par « mot ». Par contrat avec les Éditions Champion, nous sommes chargés d'élaborer un système d'étiquetage exhaustif et exact en vue de l'hypertextualisation du corpus avec moteur de recherche et cartographie statistiques.

Au niveau des formes composées, notre objectif est double. Nous souhaitons inventorier l'ensemble des récurrences propres à Balzac, élaborer un dictionnaire des formes composées du corpus de *La Comédie humaine*. Chaque auteur, chaque texte se caractérise en effet par un emploi particulier des composés et plus encore par une créativité particulière dans le riche domaine des figements. Par ailleurs, nous entendons renforcer, par l'expérience d'un vaste corpus cohérent, l'efficacité des dictionnaires de composés qui nous servent de référence, ceux du LADL et du système INTEX, coordonnés par Blandine Courtois (DELACF), ainsi, bien sûr, que les bibliothèques de transducteurs, le cas échéant.

Nous proposons tout d'abord de recenser, par un balayage exhaustif du texte, l'ensemble des séquences de *n* formes simples, puis de les filtrer suivant divers critères statistiques avant de les examiner *à l'œil et à la main* selon des critères linguistiques apparentés à ceux du LADL.

Nous travaillons sur un texte quasiment brut, à l'exception d'une normalisation par recodage des articles définis contractés (et des indéfinis *du* et *des*) en [Prep *le(s)*]. À ce stade, nous ne craignons pas les effets de l'ambiguïté ainsi créée entre des déterminants et des pronoms clitiques de même forme (dans nos fichiers de travail originaux, *du* donne en réalité *de+ le*, le signe + servant à conserver la trace de cette opération ; mais pour la procédure ici décrite, nous devons récupérer une préposition « vierge » ; c'est donc un fichier *adhoc*).

2. Cadre théorique

Un programme constitue les séquences de *n* contigus non interrompues, ni par une ponctuation, ni par un hapax (ni même par une forme d'occurrence 2). Ces séquences sont recensées et tout d'abord classées par ordre de fréquence décroissante. Les observations suivantes sont alors permises :

La séquence « brute » de 3 « mots » la plus fréquente est [et/des (de/les)], avec 1 718 occurrences, devant [il/y/a], 1 503, et [par/des (de/les)], 1 400. Dans les tableaux suivants, nous soulignons les séquences relevant d'une identification comme unité composée.

et/de/les	1 524
il/y/a	1 503
par/de/les	1 400
ne/est/pas	1 374
en/ce/moment	1 357
il/ne/y	1 229

À/le/milieu	1 163
Le/milieu/de	1 158
Le/moment/où	1 151
À/le/moment	1 057
/ne/ai	1 053

Si l'on donne un codage commun à tous les déterminants singuliers et pluriels, la matrice la plus fréquente devient [S/de/ P], qui compte 3 598 occurrences, dont 1 138 pour [un/des(de/les)], 853 pour [une/des], mais surtout 720 et 640 occurrences respectivement pour les deux déterminants composites favoris de Balzac, [un(e)/de/ces].

La plus fréquente de toutes les séquences reconnaissables comme unité composée est [il y a], puis on trouve [en/ce/moment], avec 1 357 occ. [au(à+ /le)/milieu], avec ses 1 163 occurrences, même si elle peut être un adverbe autonome, doit d'abord être pensée comme synonyme de [au(à+ /le)/milieu/de], ce que confirme la proximité de l'autre synonyme, [le/milieu/de], qui compte 1 158 occ. Vérification faite, la suite [au/milieu/de] se retrouve 1 120 fois (en comptant les 26 occ. de [au/milieu/des(quelles)]). La situation est la même pour [au/moment] et [le/moment/où], 1 151 et 1 057 occ.

Pour les séquences de 4 « mots », les 6 plus fréquentes sont pertinentes : [il/ne(n')/y/a] doit être reconnu, au moyen d'un transducteur, comme une séquence enclosant un terme simple.

à/le/milieu/de	1 092
à/le/moment/où	923
il/ne/y/a	785

ne/est/¤ce/pas	691
à/le/lieu/de	604
à/le/fond/de	517

En revanche, dès la longueur de 5 « mots », les séquences dominantes sont déjà composites, généralement par présence d'un déterminant après une préposition ou un présentatif : la première séquence

pertinente est [le(l')/un/et/le(l')/autre] qui n'a déjà plus que 140 occ. Quant au *pas* de [il/ne(n')/y/a/pas] il ne nécessite pas un transducteur puisqu'il est linéairement distinct.

il/ne/y/a/pas	248
il/ne/y/a/que	212
à/le/milieu/de/les	210
il/y/a/de/les	191
à/le/milieu/de/le	183

à/le/moment/où/le	165
ne/y/a/pas/de	146
à/le/milieu/de/la	140
le/un/et/le/autre	140

À partir de 6 constituants, les séquences pertinentes présentent des effectifs nettement moindres ; ici, la première à prendre en considération semble être [à/le/coin/de/le/feu]. Elles deviennent aussi sémantiquement plus spécifiques.

il/ne/y/a/pas/de	144
ce/que/il/y/a/de	67
il/ne/y/a/rien/de	53
à/le/coin/de/le/feu	44

il/ne/y/a/plus/de	44
Sur/le/seuil/de/la/porte	43
la/fin/de/le/mois/de	41
à/le/fond/de/le/coeur	41

Cette tendance à la diminution des effectifs se creuse aux longueurs 7 et 8, cette dernière étant la limite supérieure de validité de la procédure :

pour/la/première/fois/de/sa/vie	38
dans/les/premiers/jours/de/le/mois/de	15
vers/la/fin/de/le/mois/de	35
tout/ce/que/il/y/a/de/plus	9
tout/ce/que/il/y/a/de	31
vous/ne/savez/pas/ce/que/c'est/que	8
ce/que/il/y/a/de/plus	29
il/y/a/je/ne/sais/quoi/de	8
il/y/eut/un/moment/de/silence	22
à/le/coin/de/la/rue/de/la	6
il/ne/y/a/rien/de/plus	19
à/le/coin/de/la/rue/de/le	6

il/ne/y/a/que/de/les	18
en/siégeant/à/le/tribunal/consulaire/et/en	6
les/premiers/jours/de/le/mois/de	17
je/ne/ai/pas/besoin/de/vous/dire	6
à/le/coin/de/la/rue/de	17
siégeant/à/le/tribunal/consulaire/et/en/combattant	6
le/chef/de/la/police/de/sûreté	16
dans/les/bras/le/un/de/le/autre	6

Ces listes, même classées par ordre de fréquence décroissante, sont par leur longueur difficiles à exploiter. Il y a notamment 52 000 séquences différentes de 3 « mots » présentant une occurrence supérieure à 5 dans notre corpus.

Il faut donc trouver des filtres supplémentaires pour alléger ces listes de ce qui n'est pas susceptible de nous intéresser.

Un premier moyen, très simple, consiste à leur appliquer les ressources d'INTEX. Nous avons établi la liste des composés reconnus par INTEX (4.11) et employé cette liste comme filtre. Le résultat nous permet en effet d'éliminer certaines séquences très fréquentes, mais aussi de constater l'absence dans le DELACF de certaines séquences, et d'ouvrir le dialogue sur ce point. Par exemple, *il y a* semble absent.

Nous allons présenter plus en détail cette phase des travaux et surtout l'application aux têtes de listes des critères de compositionnalité de Gaston Gross.

3. Les critères de compositionnalité

Nous nous sommes en effet inspirés des définitions de Gaston Gross et de Max Silberztein pour fixer les critères de compositionnalité. Nous proposons dans le présent chapitre de présenter ces différents critères. La définition du figement telle que la donne le *Dictionnaire de linguistique* est la suivante :

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

3.1. Critères de compositionnalité selon GROSS 1990

Gross distingue les séquences figées (mots composés) des séquences libres relativement à leur degré de figement. Ce degré de figement se calcule par rapport à six critères syntaxiques. Si la séquence répond à au moins un des critères, elle est considérée comme libre ; en revanche, si elle ne répond à aucun critère, elle est dite figée : il s'agit alors d'un mot composé. Étant donné la prédominance des noms composés dans la classe des composés, nous avons traité cette classe en priorité. Les six critères proposés par Gross sont :

3.1.1. *L'existence d'un lien syntaxique entre les deux substantifs N1 et N2*

Une séquence de type *N de N* est considérée comme libre s'il est possible d'établir un lien syntaxique entre les deux substantifs.

Dans les exemples qui suivent la séquence « chien (N1) de Max (N2) » est libre car il existe un lien d'appartenance entre N1 et N2, alors que la séquence « pomme (N1) d'Adam (N2) » est figée car il n'existe pas de lien syntaxique entre les deux substantifs.

Le chien de Max → *Max a un chien*
La pomme d'Adam → **Adam a une pomme*

3.1.2. *La pronominalisation*

La séquence est considérée comme libre s'il est possible de pronominaliser un des deux substantifs. Dans les deux exemples qui suivent, la première séquence est libre alors que la seconde est figée.

La voiture de Luc → *sa voiture*
La maladie d'Alzheimer → **sa maladie*

3.1.3. *La rupture paradigmatique*

Dans le cas des séquences libres, il est possible de remplacer un des substantifs par un nom de la même classe distributionnelle. Cette opération n'est guère possible s'il s'agit d'une séquence figée. Dans les exemples proposés, les deux premières séquences sont libres, la troisième est figée.

La chance de (Jean + Luc + Paul...)

*Une chance de (*Luc + *Marie + *Paul...)*
*Une chance de (cocu + *dupe)*

3.1.4. *Le figement partiel*

Le figement partiel est un signe de figement total. Nous rencontrons souvent des groupes nominaux dont une partie de la séquence est figée tandis qu'une partie reste libre. Nous proposons les exemples suivants :

Luc a versé un nuage de lait. (nuage de lait est figé)
**Luc a versé un nuage.*
Luc a versé du lait.

3.1.5. *Les insertions*

Les séquences libres permettent les insertions (adjectifs, adverbes, etc.) alors que les séquences figées n'appellent pas cette transformation.

La voiture de mon frère
La voiture neuve de mon frère
La voiture de mon jeune frère

Or,

Une pomme de terre
**Une pomme pourrie de terre*

3.1.6. *Les pluriels obligatoires*

Les pluriels obligatoires sont très souvent signe de figement comme le montrent les deux exemples suivants :

Des nerfs d'acier
Les moyens du bord

3.2. Les critères de compositionnalité selon Silberztein

Afin de distinguer une séquence libre d'un nom composé (et cette définition rejoint celle de Gross), Silberztein (1993) propose de : « comparer les propriétés syntaxiques et sémantiques des séquences "candidates au statut de nom composé" à celles générales des séquences libres. Si toutes les propriétés syntaxiques et sémantiques

d'une séquence peuvent être déduites des propriétés de ses constituants, on n'a pas besoin de les décrire explicitement dans un dictionnaire : la séquence est libre. En revanche, si au moins une propriété syntaxique ou sémantique de la séquence ne peut pas être calculée, il faut associer explicitement cette séquence et ses propriétés, et donc placer cette séquence dans un dictionnaire : la séquence est figée. »

Les quatre critères (deux critères sont d'ordre sémantique et les deux autres d'ordre syntaxique) permettant d'effectuer cette distinction sont :

3.2.1. *L'atomicité sémantique*

Le sens d'une séquence libre (groupe nominal libre) peut être calculé à partir du sens de ses constituants, par exemple, *la pomme verte* = sens de *pomme* de couleur *verte.* Ce n'est pas le cas des séquences figées ; par exemple la séquence *arme blanche* peut être soit libre soit figée ce qui rend cette séquence ambiguë.

Le sens d'un nom composé n'est pas directement calculable à partir du sens de ses constituants.

3.2.2. *L'institutionnalisation de l'usage*

Certains groupes nominaux sont utilisés de façon institutionnalisée, c'est-à-dire que nous choisissons automatiquement une séquence plutôt qu'une autre. Nous disons par exemple :

Canal alimentaire plutôt que *canal de l'alimentation.*

3.2.3. *Les restrictions distributionnelles et l'analyse transformationnelle*

Une séquence est libre s'il est possible de remplacer un des constituants (de la séquence) par un mot de la même classe distributionnelle. En revanche, si nous ne pouvons pas effectuer cette opération, la séquence est figée. Dans les exemples qui suivent, la première séquence est libre, la seconde est figée.

Le chat (chien, cheval...) de Tom.
**Les choux (tomates, carottes...) de Bruxelles.*

4. Méthodologie

Nous avons donc recensé dans le corpus « Balzac » l'ensemble des séquences de n formes graphiques présentant une récurrence (cf. 2). Nous avons alors obtenu des listes que nous avons triées afin de garder les séquences qui présentaient des traits de figement. Nous avons utilisé INTEX afin de repérer les séquences reconnues par le système. Puis nous avons identifié par des signes distinctifs les séquences qui nous semblaient être figées, les composés spécifiques à l'auteur et nous les avons intégrés ou non aux différentes classes de dictionnaires, ici aux dictionnaires de corpus (notamment des expressions figées et des noms de personnage).

Nous proposons d'étudier les récurrences de 5, 6, 7 et 8 formes graphiques.

5. Recherches de récurrences dans un corpus littéraire

5.1. Récurrences de 8 mots

La première étape de notre travail consiste à trier dans les listes obtenues les séquences figées selon les critères précédemment énoncés. Nous avons donc attribué à certaines séquences des codes de reconnaissance.

A	B	C
dans les premiers jours du mois de	15	G
entre une heure et deux du matin	5	G
vers la fin du mois d'octobre	5	g
les premiers jours du mois de décembre	5	g
vers la fin du mois de mai	4	g
vers la fin du mois de février	4	g
vers la fin du mois de décembre	4	g
vers la fin du mois de septembre	4	g
les premiers jours du mois de mai	4	g
par une belle matinée du mois de	4	g
vers les derniers jours du mois de	3	g
vers la fin du mois de juillet	3	g
vers la fin du mois de mars	3	g
au commencement du mois de décembre	3	g

adjoint au maire du deuxième arrondissement	3	d
un huissier du cabinet du roi	3	d
premier gentilhomme de la chambre du roi	3	d
tout ce qu'il y a de plus	9	

Tableau 1 : liste des séquences (des récurrences de 8 mots) présentant des traits de figement.

Nous avons ainsi obtenu dans la colonne A : les séquences de X mots, dans la colonne B : le nombre d'occurrences de la séquence et dans la colonne C : nous avons sélectionné par un code simplifié (**d** pour dictionnaire, **g** pour grammaire, etc.) les séquences qui nous semblaient présenter des traits de figement.

Une fois le tri effectué, nous avons obtenu une liste de séquences que nous avons traitée comme du texte sous INTEX afin de repérer les séquences non reconnues. (Les séquences soulignées sont celles reconnues par INTEX, les séquences surlignées sont celles que nous avons retenues dans la liste car elles présentent des traits de figement).

adjoint au maire du deuxième arrondissement
un huissier du cabinet du roi
premier gentilhomme de la chambre du roi
au commencement du mois de décembre
dans les premiers jours du mois de
entre une heure et deux du matin
vers les derniers jours du mois de
vers la fin du mois de juillet
vers la fin du mois de mai
vers la fin du mois de février
vers la fin du mois de mars
vers la fin du mois de décembre
vers la fin du mois de septembre
vers la fin du mois d'octobre
les premiers jours du mois de décembre
les premiers jours du mois de mai
par une belle matinée du mois de .

Ainsi, nous avons remarqué qu'INTEX ne reconnaissait pas des séquences telles que : *Nième arrondissement, cabinet du roi, chambre du roi, etc.*

5.2. Les récurrences de 7 mots

Nous avons effectué le même travail pour les séquences de 7 mots. La liste initiale contient 890 séquences ; nous en avons retenu 36 (cf. tableau 2).

au parquet du procureur général	la fin du mois de septembre
il y a du vrai dans	les dimanches et les jours de fête
au pied du lit de	par une belle matinée du mois
dans les mystères de la vie privée	premiers jours du mois de mai
au pain et à l'eau	la fin du mois de mai
droit de vie et de mort sur	une heure et deux du matin
le chef de la police de sûreté	les derniers jours du mois de
vers les derniers jours du mois	la fin du mois de décembre
le commencement du mois de mars	premiers jours du mois de décembre
le mois d'octobre de l'année	sur les deux heures du matin
le commencement du mois de décembre	la fin du mois d'octobre
le commencement du mois d'octobre	jusqu'à deux heures du matin
onze heures et demie du soir	vers le milieu du mois de
la fin du mois de février	dans les premiers jours du mois
dix heures et demie du soir	au commencement du mois de
une belle matinée du mois de	les premiers jours du mois de
le valet de chambre du sénateur	vers la fin du mois de
	pour la première fois de sa vie
	d'un bout à l'autre de

Tableau 2 : liste des séquences (des récurrences de 7 mots) présentant des traits de figement.

Comme nous pouvons le constater dans la liste ci-dessous, INTEX a reconnu un grand nombre de séquences telles que *mois de novembre*, *chef de la police*, *etc.* mais n'a pas reconnu les séquences *mystères de la vie privée*, *droit de vie et de mort*, *pied du lit*, *police de sûreté*, *parquet du procureur*, etc., malgré le fait que ce sont des séquences figées. Nous les avons donc intégrées aux dictionnaires de corpus.

au parquet du procureur général	le commencement du mois de décembre
il y a du vrai dans	le commencement du mois d'octobre
au pied du lit de	onze heures et demie du soir
dans les mystères de la vie privée	la fin du mois de février
au pain et à l'eau	dix heures et demie du soir
droit de vie et de mort sur	une belle matinée du mois de
le chef de la police de sûreté	le valet de chambre du sénateur
vers les derniers jours du mois	la fin du mois de septembre
le commencement du mois de mars	
le mois d'octobre de l'année	

les dimanches et les jours de fête
par une belle matinée du mois
premiers jours du mois de mai
la fin du mois de mai
une heure et deux du matin
les derniers jours du mois de
la fin du mois de décembre
premiers jours du mois de décembre
sur les deux heures du matin
la fin du mois d'octobre
jusqu'à deux heures du matin
vers le milieu du mois de
dans les premiers jours du mois
au commencement du mois de
les premiers jours du mois de
vers la fin du mois de
pour la première fois de sa vie
d'un bout à l'autre de

5.3. Les récurrences de 6 mots

Nous avons relevé 4 050 occurrences de 6 mots et retenu 174 séquences présentant des traits de figement. Nous présentons ci-dessous une partie de la liste.

sur le seuil de la porte.
la fin du mois de.
au fond de son cœur.
sur le pas de la porte.
le commencement du mois de.
à deux heures du matin.
la veille du jour où.
à cinq heures du matin.
premiers jours du mois de.
Au commencement du mois.
Au commencement de l'année.
vers la fin de l'année.
à l'heure du dîner.
mille francs de lettres de change.
Au bout d'un mois.
à dix lieues à la ronde.
à sept heures du matin.
à neuf heures du matin.
jusque à deux heures du.
à la fin du mois.
deux cent mille francs de dot.
Au bout de dix ans.
fin du mois d'octobre.
L'index de sa main droite.
à trois heures du matin.
Au bout d'une demi-heure.
le premier étage de la maison.
douze cents francs de rente viagère.
un billet de cinq cents francs.
fin du mois de mai.
jours du mois de mai.
à sept heures du soir.
heure et deux du matin.
fin du mois de décembre.
jours du mois de décembre.
de cinquante mille livres de rente.
les derniers jours du mois.
à la fin de la soirée.

Nous avons relevé des séquences non reconnues par INTEX, séquences que nous avons intégrées au dictionnaire de corpus :

seuil de <DET> porte,
<Nième> étage,

en proie à,
la pointe du pied,
au bout de ...

5.4. Les récurrences de 5 mots

Nous présentons, en dernier lieu, une partie de la liste des séquences de 5 mots. Les noms composés reconnus par INTEX (soulignés) sont plus nombreux que pour les listes précédentes. Cependant, il reste des séquences non reconnues qui sont pourtant figées.

le procureur du roi
la plupart du temps
le seuil de la porte
le fond du cœur
le coin du feu
les gens du monde
au bout du
les choses de la vie
au sein de la
la maîtresse de la maison
vers la fin de le
au fond d'une
au commencement de le
par un coup d'œil
en proie à des
en un clin d'œil
les gens du peuple
le coucher du soleil
d'un coup d'œil
une femme à la mode

à l'aspect du
l'heure du dîner
le maître des cérémonies
le pied du lit
chef de la police de
la veille du jour
le reste de ses jours
la pointe des pieds
à l'abri des
au chevet du
les gens du pays
des lettres de change
le coin de l'œil
du juge de paix
le chef de la police
juge au tribunal de
la rapidité de l'éclair
les lois du monde
les gens de la maison
au tribunal de commerce

du mois d'octobre
le tribunal de première instance
au moyen d'un
au pied des
le chef de la maison
la pointe du pied
le maître du logis
les gens du château
des bas de soie
à la hauteur du
du mois d'avril
les lois de la nature
des gens d'esprit
dans sa robe de chambre
le garde des sceaux
le président du tribunal
du maître de poste
les mystères de la vie
le chemin le plus court
au mois de septembre

Les séquences qui nous semblent figées dans cette liste sont : *fond du cœur*, *pied du lit*, *maîtresse de maison*, *clin d'œil*, *gens du château*, *chambre à coucher*, séquences très fréquentes dans l'œuvre de Balzac.

6. Nos premières conclusions...

Les mots composés que nous avons retenus dans cette première recherche seront donc intégrés aux dictionnaires de corpus. Nous avons déjà repéré dans les listes étudiées des séquences non reconnues par INTEX présentant des traits de figement. Nous en avons proposé un échantillon (les séquences soulignées dans la liste ci-dessous sont celles reconnues par INTEX).

(de) la veuve et (de) l'orphelin
à <NB> lieues à la ronde
à faire pitié
à la fonte des neiges
à la sueur de mon <Adj.possessif> front.
adjoint au maire du deuxième arrondissement.
au bout de
au chevet du lit
au ministère de la guerre
Au nom du ciel.
au parquet du procureur
au pied du lit
au pied du mur
avec la rapidité de l'éclair
chambre à coucher
chambre du condamné à mort
chef de la police de sûreté
chef de la police de sûreté
dans la gueule du loup
des airs de reine
droit de vie et de mort
en proie à
entre la vie et la mort
feux du soleil couchant
fièvre nerveuse
homme de la trempe
huissier du cabinet du roi .
juge au tribunal de commerce
la pluie et le beau temps
le chemin le plus court
le premier jour de l'an
les allées et venues
maire du deuxième arrondissement
mettre les points sur les i
mystères de la vie privée
parquet du procureur général
premier gentilhomme de la chambre du roi
président du tribunal de commerce
se jeter au cou
seuil de la porte
sur la pointe du pied
sur le bout du doigt
tribunal de première instance

Cette liste n'est évidemment pas exhaustive car il nous reste à traiter les récurrences de 4, 3 et 2 mots.

Notre travail comprend également la construction d'un dictionnaire des noms composés de personnages. Nous présentons la liste à partir du roman *La Femme de trente ans* :

Comte Alfred de Vandenesse, Comte de Aiglemont, Comte de Artois, Comte de Saint-Héreen, comtesse de Listomère-Landon, Duc de Angoulême, Gustave marquis de Aiglemont, La comtesse Moïna de Saint-Héreen, M. Charles de Vandenesse, M. de Aiglemont, M. de Mauny, M. de Ronquerolles, M. de Vandenesse, M. de Wimphen, M. le baron de Mauny, M. le curé, M. le Maire, Madame de Aiglemont, Madame de Sérizy, marquis de Vandenesse, marquise de Aiglemont, Mme de Aiglemont, Mme de Listomère, Mme de Saint-Héreen, Mme de Sérizy, Mme de Wimphen, Mme la comtesse de Sérizy, Mme la comtesse de Sérizy, Monsieur de Marsay, Monsieur de Ronquerolles, Princesse de Cadignan.

Le travail que nous avons présenté est à ses prémices mais les résultats que nous avons obtenus pour les récurrences de 5, 6, 7 et 8 mots sont d'ores et déjà très encourageants.

7. Le filtrage statistique

Nous présentons maintenant une autre procédure de filtrage statistique, plus sophistiquée que la simple fréquence.

Un critère absolu de compositionnalité est bien l'exclusivité d'occurrence d'une forme simple dans une séquence et une seule. Un cas très typique dès la surface brute du texte est *d'emblée*, codé [de/emblée], dans la mesure où *emblée* n'a aucune chance d'apparaître autrement que derrière *d'*. Si on compare l'occurrence de *emblée* et de [de/emblée], on constatera normalement l'égalité. C'est la même chose, après codage des déterminants, pour *instar*. On comprendra mieux maintenant la proposition de regrouper les déterminants les plus courants, pour cette phase du travail, sous un codage unique (S ou P pour sing. ou plur.), qui permettra de reconnaître une même séquence [à/ S/instar]. En effet, l'occurrence de cette séquence sera égale à celle d'*instar*, ce qui n'est pas le cas de *à l'instar de*, en raison de la concurrence de, *i.e.*, *à son instar*.

L'idée est d'affecter toutes les séquences récurrentes du corpus d'un indice de « saturation », calculé selon le quotient de l'occurrence du moins fréquent de ses constituants sur l'occurrence de la séquence entière (en pourcentage). Plus ce pourcentage sera fort, plus les chances de reconnaître un composé ou un figement sont fortes.

Un programme génère donc les tableaux suivants :

et/de/les	1 524	19	Et
il/y/a	1 503	128	Y
par/de/les	1 400	50	Par
ne/est/pas	1 374	59	Est
en/ce/moment	1 357	325	moment
il/ne/y	1 229	105	Y
à/le/milieu	1 163	867	milieu
le/milieu/de	1 158	863	milieu
le/moment/où	1 151	276	moment
à/le/moment	1 057	253	moment

Les deux premières colonnes reprennent les tableaux fréquentiels déjà évoqués, la quatrième contient le constituant le moins fréquent de la séquence, et la troisième le quotient (pour 1 000) de l'occurrence de ce dernier sur celle de la séquence.

On aura évidemment intérêt à reclasser ces listes en combinant un ordre décroissant d'indice pour une classe déterminée de fréquence, ce qui permet de mettre en évidence, pour les séquences de 3 « mots » ayant plus de 100 occurrences dans le corpus, la « tête de liste » suivante :

tous/les/soirs	132	978	soirs
heures/et/demie	200	957	demie
la/plupart/de	309	912	plupart
se/adressant/à	159	893	adressant
à/le/delà	177	885	delà
le/seuil/de	119	875	seuil
à/le/milieu	1 163	867	milieu
le/milieu/de	1 158	863	milieu
en/se/adressant	148	831	adressant
ces/sortes/de	152	822	sortes
à/voix/basse	323	788	basse
le/bord/de	211	787	bord
le/commencement/de	146	756	commencement
le/fond/de	668	677	fond
le/delà/de	135	675	delà
il/se/agit	333	665	agit
il/se/agissait	162	653	agissait

valet/de/chambre	264	652	valet
le/sein/de	254	645	sein
/le/oreille	525	643	oreille

Et aussi une « queue de liste » dont on mesure bien l'absence radicale d'intérêt !

et/à/la	183	2	à
/que/le	156	2	que
que/à/la	135	2	que
ne/se/en	112	2	se
de/une/de	110	2	une
à/le/un	109	2	un
que/il/le	105	2	il
/de/le	118	1	de

On constate ainsi que sur les 135 occurrences de *soirs*, 132 sont impliquées dans la séquence *tous les soirs*. Les seules occurrences libres sont :

l'espérance des soirs empourprés
les longs soirs d'hiver
par les soirs d'hiver

pour un ensemble de 4 millions de mots ! Situation difficilement prévisible intuitivement...

On constate par ailleurs l'extrême figement balzacien autour de l'item *adressant*, que confirme mieux encore l'examen des séquences de 4 « mots » :

en/se/adressant/à	145	815	adressant
à/le/milieu/de	1 092	814	milieu
à/le/delà/de	133	665	delà
la/plupart/de/les	208	614	plupart
il/se/agissait/de	132	532	agissait
à/le/fond/de	517	524	fond
à/le/lieu/de	604	509	Lieu
à/le/bout/de	299	509	bout
il/se/agit/de	253	505	Agit
à/le/sein/de	193	490	sein

D'une manière générale, on repère ici deux ordres de faits : des composés « de langue », comme *au milieu de*, *la plupart des*, et des spécificités de corpus, comme *en s'adressant à*.

C'est encore plus net si l'on abaisse le seuil de fréquence absolue de la séquence (ici, à 10).

ne/tarda/pas	25	1000	tarda
ne/tarda/pas/à	24	960	tarda
un/clin/de/oeil	24	1000	clin
/clin/de	24	1000	clin
clin/de/oeil	24	1000	clin
en/hochant/la/tête	37	925	hochant
à/le/improviste	20	1000	improviste
maître/de/les/requêtes	70	909	requêtes
le/instar/de	18	1000	instar
à/le/instar/de	16	889	instar
lit/de/sangle	16	1000	sangle
en/se/efforçant/de	12	857	efforçant
une/infinité/de	16	1000	infinité
en/un/clin/de	20	833	clin
les/marais/salants	16	1000	salants
de/une/voix/tonnante	15	833	tonnante
le/énormité/de	15	1000	énormité
en/se/adressant/à	145	815	adressant
se/efforçant/de	14	1000	efforçant
à/le/milieu/de	1092	814	milieu

Conclusion

En conclusion, nous voudrions évoquer le module « mots composés » de la procédure générale d'étiquetage DIATAG que nous cherchons à développer.

Cette procédure se voulant exhaustive et rationnelle, devant être appliquée à des corpus littéraires en vue de leur hypertextualisation, elle comporte nécessairement, pour chacune des passes successives à envisager, une convivialité ergonomisée.

L'étiquetage des composés devra être une phase très précoce, constituée de passes du plus long vers le plus court. Nous traiterons

d'abord les séquences de 10 constituants, puis de 9, etc. Ainsi aurons-nous intérêt à connaître l'indice de « saturation » des séquences fréquentes. Ainsi voyons-nous que *en un clin d'œil* n'est pas saturé à 100 % :

en/un/clin/de/oeil	20	833	clin
en/se/frottant/les/mains	17	708	frottant
le/maître/de/les/cérémonies	21	457	cérémonies
le/garde/de/les/sceaux	12	429	sceaux
sur/le/seuil/de/la	57	419	seuil
en/coupant/la/parole/à	12	414	coupant
le/seuil/de/la/porte	53	390	seuil
le/tribunal/de/première/instance	14	368	instance
le/maître/de/les/requêtes	28	364	requêtes
Dans/le/embrasure/de/la	29	349	embrasure

Les autres occurrences de *un clin d'œil* sont :

Un clin d'œil a décidé de ma vie.
C'est effrayant ,c'est la vie ou la mort , dans un clin d'œil.
Voir tout cela de un clin d'œil [...]

Bibliographie

GROSS (Gaston) : 1996, *Les Expressions figées en français : noms composés et autres locutions*, (Paris : Ophrys, Collection « L'Essentiel français »).

GROSS (Gaston) : 1990, « Définitions des noms composés dans un lexique-grammaire », *Langue Française*, 87 (Paris : Larousse), p. 84–90.

SILBERZTEIN (Max) : 1990, « Le dictionnaire électronique des mots composés », *Dictionnaires Électroniques du Français*, *Langue française*, 87 (Paris : Larousse), p. 71–84.

SILBERZTEIN (Max) : 1993, *Dictionnaires électroniques et analyse automatique de textes, le système INTEX* (Paris : Masson).

VIPREY (Jean-Marie) : 1997, *La Dynamique des Fleurs du Mal* (Paris : Masson).