CORPUS ET DIACHRONIE : DE LA CONSTITUTION AU TRAITEMENT
Un cas d'espèce : le roman sentimental moderne

Magali BIGEY
LASELDI, Université de Franche-Comté

SOMMAIRE

Résumé : *La constitution d'un corpus de littérature sérielle pose différents problèmes, dont ceux du recueil de données et des droits de reproduction. Une fois ces problèmes écartés, se pose alors la question du traitement. Quel type appliquer au corpus, dans quel but, quels résultats peut-on attendre ?*
Ainsi, après une brève présentation du thème de recherche, nous évoquerons les traitements effectués (création d'AFC avec le logiciel ASTARTEX, création de dictionnaires électroniques spécifiques à l'aide du logiciel Nooj) sur ce corpus réunissant 50 romans numérisés, publiés de 1978 à 2004.

1. Roman sentimental et littérature sérielle

Ce travail porte sur l'évolution du roman sentimental depuis 1942. Le corpus est divisé en deux parties, constituées de romans sentimentaux de type sériel. La première partie du corpus est le support d'une analyse narratologique[1], la seconde d'une analyse lexicologique, le tout en diachronie.
L'objet qui nous intéresse aujourd'hui est la seconde partie du corpus, qui est la partie numérisée.
Le roman sentimental moderne appartient à la littérature sérielle. Son apparition en France remonte à 1978, avec l'arrivée du roman Harlequin, parangon du genre encore aujourd'hui.
Cette littérature voit des éditions multiples, qui ne restent sur le marché que deux ou trois semaines en moyenne, avant d'être retirées de la vente et détruites, pour être aussitôt remplacées par d'autres.
Ils sont tous (ou presque) traduits de l'anglais, mais la traduction n'est pas l'objet de ce travail. Nous avons construit notre corpus à partir des traductions françaises des romans.

2. Le corpus

2.1. Du corpus papier…

[1] Le schéma narratif canonique de référence est celui dégagé par Julia Bettinotti et son équipe.

Pour la constitution de ce corpus diachronique, la première difficulté a été de trouver des romans parfois disparus depuis plusieurs décennies.
Le recours aux petites annonces et autres foires aux livres et brocantes a permis de réunir plusieurs centaines de romans, publiés de 1978 à 2004.
Ce type de démarche est un passage obligé pour toute recherche en littérature sérielle.
À la suite de cette collecte, cinquante romans ont été retenus. Le principal critère de sélection, pour les romans datant d'avant 1990, est leur date de parution.
Le corpus réuni couvre une période de 26 ans.
Le résultat est un corpus de 8000 pages papier, très peu exploitable. C'est pour cette raison que nous avons eu recours à la numérisation.

2.2. Au corpus numérisé

Afin d'avoir un corpus complet, il a fallu numériser chaque page de chaque roman.
Un logiciel de reconnaissance automatique de caractères (OCR) a permis d'obtenir un corpus presque exploitable et analysable par des logiciels de traitement automatique de textes.
Le corpus final de 2 millions de mots représente 3146 pages Word[1].

Nettoyage du corpus

L'OCR n'a pas donné un texte exploitable immédiatement, et la phase de préparation a été encore longue. Un tel corpus doit être nettoyé de ses « coquilles » et « mots inconnus » afin de pouvoir être utilisé.

Corrections

La première étape de correction s'est faite sur écran. Chaque page numérisée est relue et débarrassée des éléments « inconnus » dus à l'océrisation, mais beaucoup d'entre eux passent à côté de la vigilance du lecteur.
Il reste encore beaucoup d'éléments non reconnus, identifiés grâce au logiciel Nooj[2]. Est alors créée une liste des mots inconnus.
Ces mots inconnus ont diverses origines et doivent être traités de manière différente.
Certains pourront être corrigés automatiquement, d'autres nécessiteront des traitements au cas par cas.

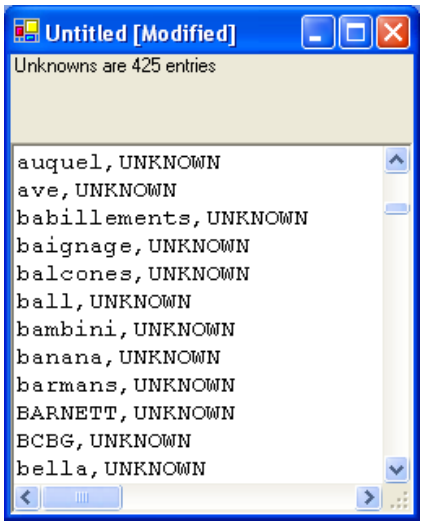

Fig. 1a : Liste des mots inconnus, éditée par le logiciel.

Les mots inconnus sont de plusieurs types :
- les mots inconnus de type *« coquilles »*, soit dus à la numérisation (le logiciel de reconnaissance automatique reconnaît souvent un *« rn »* pour un *« m »*, un *« sreur »* pour *« sœur »*, un *« 1 »* pour *« l »* ex : *panta1on*) et qui ont échappé à la correction manuelle, soit des *« coquilles »* présentes dans l'exemplaire papier du roman. (ex : *« baignage »* pour *« baignade »*, *« corgase »* pour *« corsage »*).

[1] Pages format Word, police Times New Roman, 12 points.
[2] Ce logiciel de traitement automatique est développé par Max Silberztein, laboratoire LASELDI, université de Franche-Comté.

- problème de découpage des mots par le logiciel. On trouvera par exemple *« er »* considéré comme une erreur, alors qu'il appartient au mot *« amer »*.
Pour la correction des coquilles du type *« sreur »*, nous avons utilisé la fonction « rechercher-remplacer » de Word, qui permet une correction automatique des différentes occurrences.
- les mots inconnus qui sont en fait des mots étrangers (anglais, espagnols, tibétains…).
Ils devront être traités à part, mais il est essentiel de les garder intacts dans le texte.
- les mots inconnus qui sont des « mots d'enfants », tel *« éphélan »* pour éléphant, ou encore des mots « inventés » par le traducteur, sortes de néologismes ou d'abus de langage tels *« funambuliste »* ou *« crispement »*[1].
- les mots inconnus qui correspondent à des retranscriptions d'hésitations dans des dialogues : tels : *« abso… »* pour *« absolument »*, *« tre »* pour *« en-tre »*.
- les mots inconnus des dictionnaires électroniques mais pas du TLFi[2] tels *« babillements »* ou *« poincianas »* (qui est une plante).

Certaines de ces formes sont intégrées à un dictionnaire filtre[3] afin d'être reconnues par le logiciel. Le dictionnaire est spécifique à ce corpus. Il comprendra les mots étrangers, les noms propres non reconnus, les mots inconnus des différents dictionnaires, les abus de langage et élisions trouvées dans les dialogues. Ces mots sont importants pour l'analyse et doivent rester en l'état.
D'autres sont corrigées. Ce sont celles qui résultent de la numérisation, les coquilles dues au scanner.

3. Le traitement automatique

Parmi toutes les possibilités de traitement offertes par les logiciels de traitement automatique, il faut choisir la méthode adaptée au type de travail souhaité.
Plusieurs possibilités s'offrent à nous :

3.1. La lemmatisation

Très efficace pour les analyses de champs sémantiques, elle regroupe sous un même lemme des formes graphiques différentes. Malheureusement, ce type de traitement entraîne une perte d'information sémantique :
Exemple : *vouloir*, *veux*, *voudrais*, *aurais voulu*… ont des sens très différents, et seraient pourtant réunis sous une même « étiquette ».

3.2. L'étiquetage

Cette méthode consiste à donner à des formes graphiques une étiquette relevant de leur statut. Il existe les étiquetages morphologiques et morphosyntaxiques.
Un des avantages de l'étiquetage est qu'il permet de faire des recherches très précises du type : *« tous les verbes au passé simple »*, *« toutes les occurrences du verbe aimer au présent »* ou *« tous les noms de parties du corps »…*

3.3. L'Analyse Factorielle des Correspondances (A.F.C.)

> « L'analyse factorielle traite des tableaux de nombres et elle remplace un tableau difficile à lire par un tableau plus simple à lire qui soit une bonne approximation de celui-ci. »[4]

L'analyse factorielle donne une « cartographie » de la répartition du vocabulaire dans le corpus et isole des phénomènes qui seraient peut-être passés inaperçus.
En effet, il est difficile de visualiser la différence entre deux extraits de textes qui paraissent identiques dans leur style et leur vocabulaire :

---

[1] Ces mots « inventés » sont quasi inexistants depuis la fin des années 80. Les traducteurs sont aujourd'hui des personnes qui ont fait de hautes études littéraires et qui parlent parfaitement l'anglais.
[2] Trésor de la Langue Française informatisé, http://atilf.atilf.fr
[3] Ce dictionnaire est ensuite intégré au logiciel Nooj.
[4] Cette citation est tirée du Que sais-je ? de Philippe Cibois, *L'analyse factorielle : analyse en composantes principales et analyse des correspondances*, p. 5.

> Exemples :
> Une légère brise agita les voilages de mousseline. Dehors, les criquets, cachés dans l'herbe haute, firent entendre leur cri strident et monotone. La lumière dorée du couchant filtrait à travers les tentures. Cette soirée de printemps ressemblait à des milliers d'autres, et pourtant je savais qu'elle était différente. Elle marquait la fin d'une époque.[1]
> Si l'on se fiait aux brochures, l'île ne mesurait pas plus de huit kilomètres de long. Mais les routes avaient été tracées de manière à respecter au mieux la forêt, si bien qu'il fallut zigzaguer pendant un quart d'heure avant d'atteindre le parking de l'auberge Seagrass. Un portier en livrée bâillait copieusement sur le seuil.[2]

Les extraits ci-dessus sont issus de deux textes, numérotés 17 et 23 sur la représentation suivante.

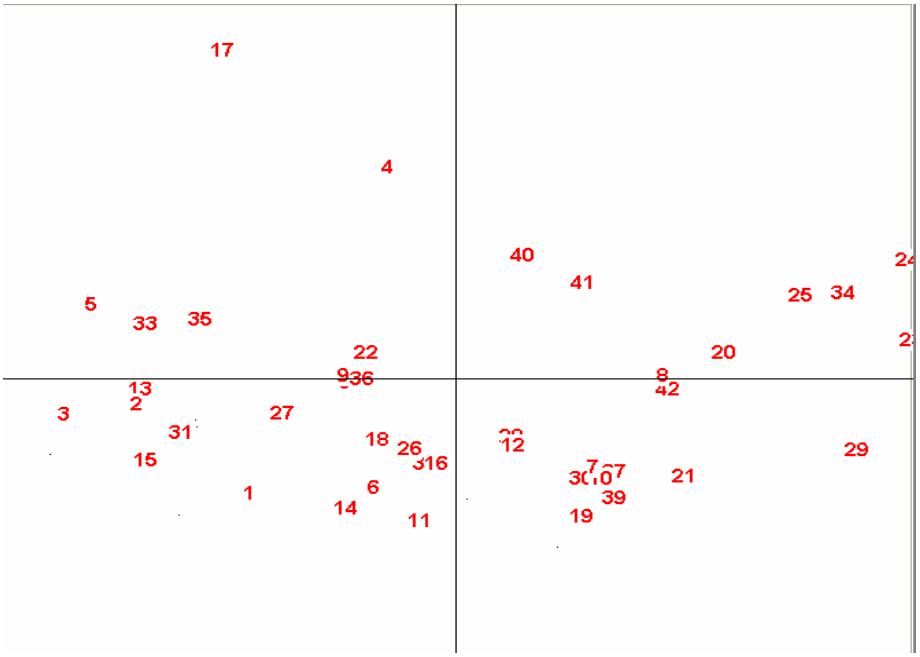

Fig. 1b : AFC de la répartition des romans en fonction de leur vocabulaire[3]

Un simple regard sur le résultat de l'AFC donne des directions de recherche. Ici, plusieurs hypothèses s'offrent à nous.
Après vérification, il s'est avéré que le texte 17 est le seul du corpus à avoir une orientation policière, et on peut supposer que c'est ce qui explique sa situation dans le tableau d'analyse. Mais en y regardant de plus près, on peut voir que ce texte 17 a un narrateur homodiégétique, ce qui est très rare dans le roman sentimental sériel. Ces romans sont très souvent constitués essentiellement de dialogues et le narrateur est la plupart du temps hétérodiégétique, ce qui est le cas pour le texte 23.
Une autre indication est donnée par le groupe formé à droite de l'AFC. Il s'avère après vérification que les romans 20, 24, 25 et 29 ont tous été écrits par le même auteur, Penny Jordan.
Cette indication oriente une nouvelle piste de recherche.

3.4. Le choix final

Nous avons finalement choisi de travailler sur un texte non lemmatisé, pour limiter la perte d'information sémantique, mais ce choix augmente considérablement le nombre des hapax dans la liste de vocabulaire.
Nous travaillons aussi sur une version du texte étiqueté morphologiquement, de manière automatique (par le logiciel Nooj). Nous avons décidé d'ignorer pour l'instant la marge d'« erreur »

[1] Rebecca Flanders, 1994, *« Vertigo »*, Sixième Sens, Harlequin, Paris, p.6.
[2] Regan Forest, 1990, *« La maison du cauchemar »*, Suspense, Harlequin, Paris, p.1.
[3] Cette AFC est issue du logiciel ASTARTEX, développé par Jean-Marie Viprey, laboratoire LASELDI, Université de Franche-Comté.

d'étiquetage, pour une partie du travail qui consiste en une recherche des variations sociologiques et linguistiques du vocabulaire en diachronie.

4. L'analyse par dictionnaires électroniques spécifiques

La création de dictionnaires électroniques spécifiques permet d'adapter l'outil au corpus à traiter. Cela permet aussi des recherches d'occurrences et de co-occurrences plus précises, car les éléments sont codés individuellement.

4.1. Le vocabulaire des parties du corps

La fouille de ce corpus a fait émerger des tendances d'utilisation de vocabulaire. Après quelques temps, il a paru évident que le vocabulaire des parties du corps devait faire l'objet d'une étude à part entière.

4.2. Préparation du dictionnaire

Les dictionnaires sont créés pour être intégrés au logiciel Nooj.
La liste du lexique des parties du corps a été réalisée à partir de la liste totale du lexique du corpus.

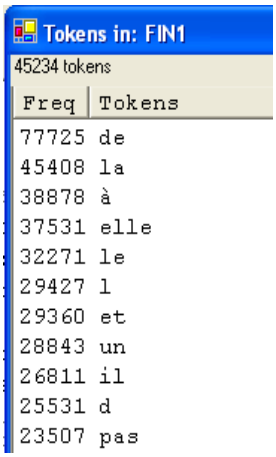

Fig.2 : Extrait de la liste des formes du corpus

Il est possible de classer le lexique par ordre alphabétique ou par nombre d'occurrences.
Nous pouvons voir ici que le texte comporte 45234 formes différentes.
Nous avons passé en revue environ 36000 occurrences (nous avons laissé de côté les hapax), afin de ne retenir que les termes utilisés pour désigner les parties du corps. Nous avons aussi laissé de côté les usages métaphoriques de certains termes, qui pourraient faire l'objet d'un autre travail.
Au final, une liste de 221 entrées a été faite. Elle constitue la base du dictionnaire électronique.
Nous tenons à signaler ici l'importance de travailler sur un texte non lemmatisé. Nous traitons par deux entrées, par exemple *« rein »* et *« reins »*.
Il est important de les différencier, car dans ce cas, l'utilisation de *« rein »* est réservée au domaine de la maladie (greffe de rein) et de la douleur[1], alors que l'utilisation de *« reins »* est réservée à la description du corps avec *« chute de reins »*, *« creux des reins »*… Dans ce deuxième cas, seules deux occurrences sur 46 font référence au domaine médical.

4.3. Création du dictionnaire

Une fois la liste du lexique des parties du corps établie, il faut coder le dictionnaire. Chaque occurrence doit pouvoir être reconnue comme *« Partie du corps »* par le logiciel.
Les codes sont propres à chaque dictionnaire et il suffit d'inventer son propre code en fonction de ce qu'on souhaite rechercher.
Le premier code à appliquer est le code « Parties du corps », afin que chaque terme soit reconnu comme tel. Le code choisi est PdC.
Ensuite, il faut ajouter d'autres codes.

[1] Dans le corpus, sur 36 occurrences de *« rein »*, une seule n'est pas rattachée au domaine médical.

Ici, les choix suivants ont été faits :
+N : nom
+f : féminin
+m : masculin
+s : singulier
+p : pluriel
+sup : partie supérieure du corps
+inf : partie inférieure du corps
+int : à l'intérieur du corps
+mem : partie des membres
+vis : partie du visage
+tet : partie de la tête
+y : yeux
+sex : sexuel
Il est très important de surcoder un dictionnaire, afin de pourvoir faire des concordances sur des demandes bien spécifiques.
Si le dictionnaire codé ne présentait que « PdC », avec genre et nombre, comme on pourrait s'y attendre, il serait impossible de faire une recherche sur tous les mots du lexique qui décrivent la tête, ou le bas du corps, ou encore le champ sémantique descriptif des yeux…

Extrait du dictionnaire :
annulaire,N+m+s+sup+mem+PdC
articulations,N+f+p+int+PdC (…)
chevelure,N+f+s+sup+tet+PdC
cheveu,N+m+s+sup+tet+PdC
cheveux,N+m+p+sup+tet+PdC (…)
nombril,N+m+s+inf+sex+PdC (…)
nuque,N+f+s+sup+PdC (…)
œil,N+m+s+sup+tet+vis+y+PdC

Une fois la liste établie dans un fichier Word, il suffit de la copier et de l'insérer dans un fichier spécifique. Après compilation, le dictionnaire est utilisable.

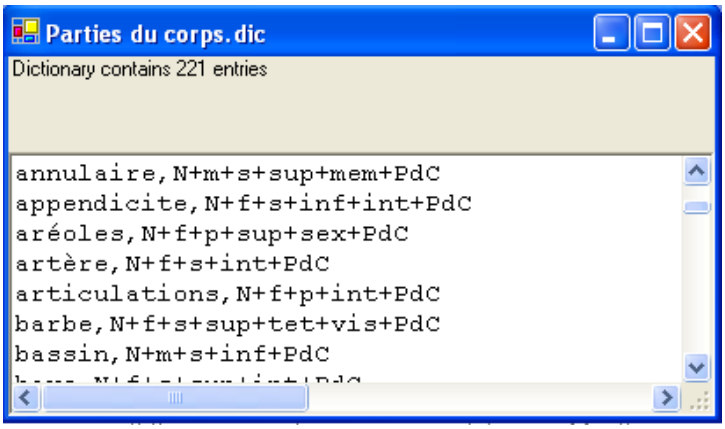

Fig. 3 : Extrait du dictionnaire

Par exemple, nous souhaitons travailler sur tous les noms féminins qui décrivent le haut du corps : Il suffit d'effectuer une recherche sur les codes N,PdC, f et sup.

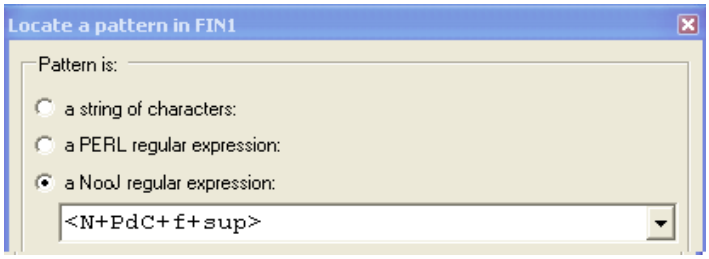

Fig. 4 : fenêtre de recherche des codes

Résultat : une liste de toutes les occurrences, en contexte, des noms féminins qui se situent dans la partie supérieure du corps.

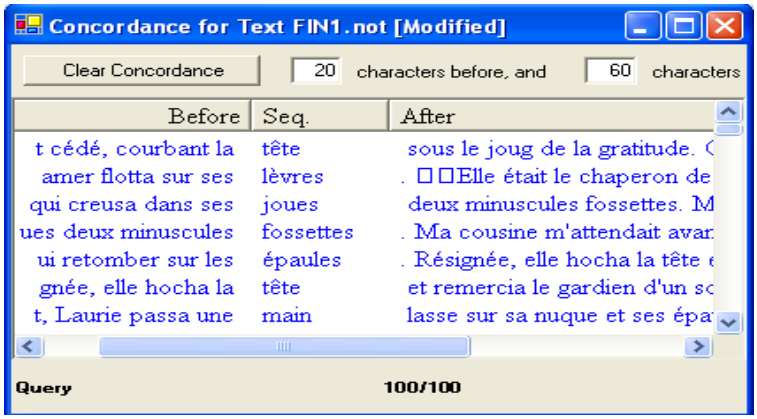

Fig. 5 : résultat de la recherche

Nous tenons à signaler que la recherche d'occurrences respecte l'ordre du texte, et présente donc les résultats en diachronie.

5. Variations en diachronie

Un corpus diachronique permet de repérer des variations, qui peuvent être de plusieurs ordres : linguistiques, sociologiques ou autres.
Un petit module statistique permet de visualiser sous forme de graphique les fréquences de chaque mot recherché.

5.1. Variations sociologiques

Exemple 1 :

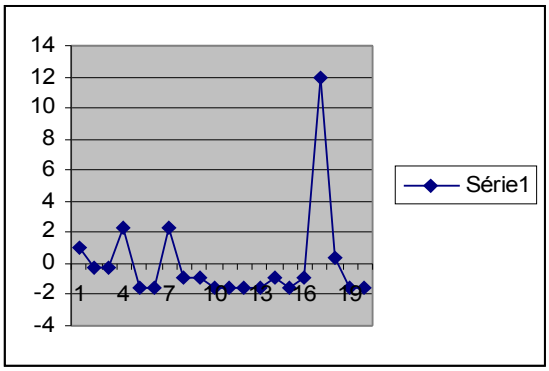

Fig. 6 : courbe représentative de l'évolution de « cigarette »

Dans cet exemple, nous remarquons la baisse d'utilisation du mot *« cigarette »* depuis les années 90, avec un seul pic important. Ce pic correspond à un roman dont le héros est dépressif, et fume beaucoup.

Exemple 2 :

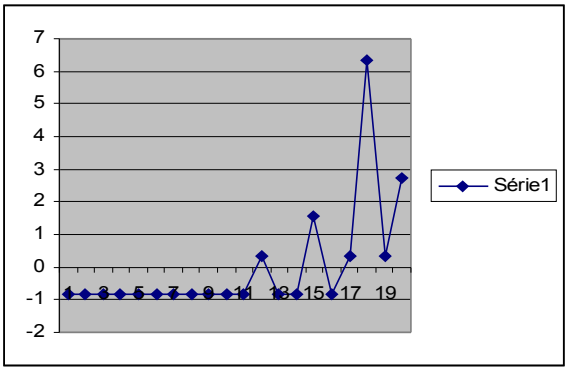

Fig. 7 : courbe représentative de l'évolution du syntagme « ordinateur portable »

Le roman sentimental appartient au genre populaire, et suit de près l'évolution de la société.
Ce phénomène explique l'importance croissante que prend le syntagme *« ordinateur portable »* à partir de la fin des années 90.

5.2. Variations linguistiques

Exemple :

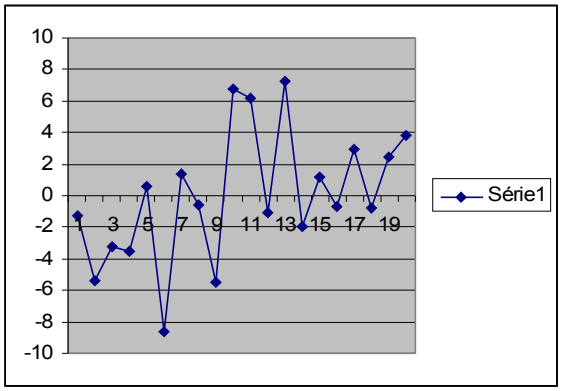

Fig. 8 : courbe représentative de l'évolution des verbes au participe passé

Une tendance d'utilisation très peu régulière se dessine. Il faut alors rechercher les occurrences en contexte afin d'en identifier la ou les causes.

Conclusion

Nous avons essayé de montrer que le choix d'un type de traitement automatique n'était pas anodin, ni forcément évident.
Parmi les différentes méthodes proposées, il en est certaines qui conviennent mieux que d'autres à l'analyse diachronique. Ici, nous avons privilégié l'étiquetage au détriment de la lemmatisation, car garder la charge sémantique des mots nous est apparu comme primordial.
Les choix de traitements vont permettre au chercheur d'orienter son travail. En quelques secondes, il lui est possible de vérifier une hypothèse, en créant par exemple une courbe représentative de l'évolution d'une forme ou d'un syntagme. Il peut alors visualiser immédiatement si l'entreprise d'une recherche est judicieuse ou non.
Le travail sur un tel corpus nécessite un traitement pluridisciplinaire. Ici, sont convoqués la linguistique-informatique, la littérature, la sociologie et le Traitement Automatique des Langues (TAL). L'important est de pouvoir faire cohabiter ces disciplines, et le traitement automatique le permet en créant un lien qui les rapproche de l'objet d'étude et dans l'objet d'étude.

BIBLIOGRAPHIE

BERNARD, M. 1999. *Introduction aux études littéraires assistées par ordinateur*, Paris, Presses Universitaires de France.

BETTINOTTI, J., (sous la dir. de). 1986, *La corrida de l'amour : le roman Harlequin*, éd. *Les cahiers du département d'études littéraires*, n°6, Montréal.
BRUNET, E. 1981. *Le vocabulaire français de 1789 à nos jours*, Genève Paris, Slatkine-Champion.
CIBOIS, P. 1983. *L'analyse factorielle : analyse en composantes principales et analyse des correspondances*, Paris, Presses Universitaires de France, Que sais-je ?.
COUEGNAS, D. 1992. *Introduction à la paralittérature*, Paris, Seuil.
HABERT, B., NAZARENKO, A. et SALEM, A. 1997. *Les linguistiques de corpus*, Paris, Armand Colin, Masson.
PEQUIGNOT, B. 1991. *La relation amoureuse, analyse sociologique du roman sentimental moderne*, Paris, L'Harmattan.
SILBERSTEIN, M. 1993. *Dictionnaires électroniques et analyses automatiques de textes : le système INTEX*, Paris, Masson.
RASTIER, F. 2001. *Arts et Sciences du texte*, Paris, Presses Universitaires de France.
VIPREY, J.-M. 2002. *Analyses textuelles et hypertextuelles des Fleurs du mal*, Paris, Champion.