Hypertexte de corpus littéraire : cartographie et statistique multidimensionnelle.

Jean-Marie VIPREY

(GRELIS, Université de Franche-Comté, F25030 Besançon)

Abstract

Literary corpuses hypertext is bridled for lack of viewpoint tools, compatible with the huge complexity and polysemy of the objects it is supposed to give access to. We intend to apply endogenous multidimensional statistics of vocabulary, and the graphs produced by them, with the aim of an open, interactive and dynamic hypertext cartography, which should match its object : pattern as well as memory of the critical browsings done and to be done. It also demands an exhaustive, previously-made and partially convivial tagging.

Résumé

L'hypertexte de corpus littéraire est bridé par sa carence en outils d'orientation compatibles avec la très grande complexité et la polysémie des objets qu'il doit viabiliser, ainsi que des parcours induits. Nous proposons d'appliquer la statistique multidimensionnelle endogène du vocabulaire, et les graphes qu'elle produit, en vue d'une cartographie hypertexte ouverte, interactive et dynamique, pertinente à son objet, modèle et mémoire des parcours critiques effectués et à faire. Cela exige aussi un étiquetage exhaustif, préalable et partiellement convivial.

Cet exposé fait suite et s'articule directement à la communication présentée lors des JADT 1998, *Une norme endogène pour le calcul stylistique du vocabulaire*, dont je rappelle les grandes lignes : on s'inscrit dans le mouvement de la statistique lexicale vers la prise en charge de la dynamique endogène des textes. On entend ici par *texte* un régime particulier des ensembles verbaux, qui active au maximum leurs aspects de densité, de complexité et de polysémie. Dans cette optique, tout niveau de structuration textuelle, et par excellence le vocabulaire, doit être modélisé comme réseau complexe, engrenage d'occurrences, macrostructure de micro-structures.

Même s'ils ont une origine et des principes communs, les outils statistiques ne reçoivent pas la même application et doivent être régulés selon que l'on s'adresse au corpus d'un point de vue documentaire ou d'un point de vue textualisant. L'analyse multidimensionnelle de grands tableaux de cooccurrence (tableaux carrés de $200<n<500$ de côté, tableaux verticaux de cette même largeur), notamment par l'A.F.C., livre des graphes en 2 ou 3 dimensions, qui synthétisent une part significative de l'information pertinente, et sont ergonomiquement lisibles comme critères de retour au texte.

Un concept-clé est celui d'*isotropie*, configuration lexicale d'ordre endogène, repérée par la parenté des profils associatifs de ses constituants, et exprimée notamment par la ventilation des items sur les graphes d'analyses factorielles. Les groupements ont l'avantage d'apparaître dans cet environnement, et contrairement aux méthodes de classification hiérarchiques, dans un continuum très propice à la saisie oculo-manuelle et au click hypertexte. De plus, les positions les moins significatives (proches des origines) sont directement repérables.

Le caractère endogène de ces organisations permet d'offrir aux chercheurs des hypothèses moins dépendantes que d'ordinaire de leurs attentes interprétatives préexistantes; notamment,

il remet en cause une vision traditionnelle figée et projective des champs lexicaux et des isotopies, au profit d'une extraction par les lignes de force du texte lui-même, de secteurs pertinents du vocabulaire, ouvrant sur des pistes thématiques entièrement renouvelées et profondément dynamiques (de proche en proche et à partir d'un seul vocable de départ - noyau du thème), on peut construire un réseau riche et complet constamment disponible pour l'extraction des contextes).
La prise en charge par niveaux du cotexte fin des occurrences dans le cadre hyperdimensionnel, permet aussi de dépasser la description éclatée et techniciste des niveaux non-lexicaux, qui caractérise l'analyse littéraire informée par la micro-linguistique.
En ce sens, la statistique multidimensionnelle favorise l'orientation objective des retours au texte, par concordances dynamiques; c'est une cartographie, qui tend à combiner au mieux l'information pertinente sur la macro et la micro-structuration. Elle offre des vues d'ensemble sur les faits microstructurels les plus saillants à l'échelle de la macrostructure.

On est donc amené à rapprocher cette proposition de l'état présent des hypertextes de corpus, notamment (mais pas exclusivement) « littéraires ». Nous entendrons par hypertextes de corpus des environnements destinés principalement non à l'extraction d'information documentaire (à visée exogène), mais à l'analyse des caractères de ce corpus (macro et/ou micro) et/ou à l'épreuve critique des champs théoriques que favorise l'expérimentation dans de vastes ensembles.
Si l'hypertexte de corpus est destiné à offrir au chercheur une viabilisation critique, ce qui caractérise l'état de l'art en ce domaine, à la différence des hypertextes documentaires (encyclopédiques), c'est l'absence flagrante d'un appareil d'orientation pertinent à son objet, c'est-à-dire tenant compte des propriétés particulières, et des textes, et de la lecture de textes.
Les produits existants proposent essentiellement de la recherche d'occurrences, à partir de saisies clavier ou de clicks plein-texte, qui reposent au mieux sur des hypothèses de travail pré-établies, qui ne peuvent s'exprimer et se tester que d'une manière morcelée à l'extrême. Une telle navigation est rendue pénible et peu stimulante, par l'absence de toute vision d'ensemble.
Paradoxalement, la plasticité qui fait l'intérêt de l'hypertexte cède ici la place à une rigidité mécanique, rapidement décourageante.
Nous proposons donc d'employer les graphes d'AFC (et, à terme, d'autres méthodes d'analyse multidimensionnelle, pourvu que les graphes produits aient au moins les mêmes qualités de continuum sur le plan lisible), aux fins d'une cartographie de l'hypertexte. La position d'un item ou d'un groupe d'items (lexical, ou d'un autre plan d'organisation) sur un graphe d'analyse multidimensionnelle, et la disposition générale de ce graphe, permet d'orienter le retour au texte (en fait, vers un contexte ou une concordance) dans la continuité d'une démarche de lecture. Le filtrage des données, que nécessitent par nature les contraintes d'affichage, se fait dans les meilleures conditions de respect de la dynamique textuelle.
L'accès aux cartes se fait par diverses voies : demande directe, conception d'une nouvelle carte ou d'un jeu de cartes à partir d'une hypothèse (liste d'items, par exemple un « champ notionnel », une liste de noms de personnages, etc), click plein texte, etc

L'hypertexte, conçu comme un environnement et non comme un objet fini, comporte ainsi dans son infrastructure, à côté de certains des outils de recherche traditionnels et d'outils statistiques déjà employés comme les spécificités et les analyses diachroniques, un nouvel appareillage destiné à sa navigation spécifique dans la textualité : des couches de texte formalisées et normalisées, pré-étiquetées par niveaux d'analyse linguistique et micro-

linguistique, un ensemble de formats de cartes, et un logiciel d'analyse multidimensionnelle en temps réel.

En effet, on doit comprendre que la cartographie dont il est ici question n'est en aucun cas figée. Certes, il existe des cartes préparées à l'avance : par exemple, celles qui organisent, selon une dimension de contexte cooccurrentiel fixée a priori (la phrase, par exemple), les classes de fréquence (les 500 premiers vocables) et/ou morpho-lexicales (les 200 premiers adjectifs, croisés entre eux ou avec les 200 premiers substantifs, etc). Cela, un atlas papier pourrait encore le faire. Mais surtout, l'environnement hypertexte ainsi conçu permet de formaliser, de visualiser, et de rendre cartographiquement fonctionnelle, toute hypothèse de travail élaborée et modifiée en cours de navigation. Par exemple, on peut projeter sur le plan la cooccurrence, avec la classe des vocables les plus fréquents, d'une liste de vocables exprimant une hypothèse thématique, et éprouver cette dernière selon une palette de modalités étendue.

De même, il existe certes, par nécessité, un ensemble de couches de texte pré-étiquetées. La « lemmatisation » en temps réel, opérationnelle pour la recherche documentaire, est en effet hors de propos si l'on veut garantir le retour au texte dans un vaste corpus. Les erreurs ou les lacunes des étiqueteurs standard dans le traitement des ambiguïtés difficiles ou irréductibles, sont ici évitées par une convivialité exhaustive, qu'un programme particulier régule. L'étiquetage (lemme, catégorie, flexion) est donc une phase préalable, à part entière. Mais, en contrepartie, la palette des couches peut être à tout instant et à la demande enrichie, suivant un codage particulier exigé par une hypothèse en cours (et ce codage sera, lui-même, achevé par convivialité). Ainsi l'automatisme est-il assigné à sa place propre, celle d'une phase dans le dialogue des hypothèses et des résultats.

Pourvu d'un historique lui-même régi par des principes statistiques, un tel environnement doit permettre de faire à l'hypertexte de corpus à finalité non-documentaire, la place qui lui revient dans les grands outils d'analyse textuelle.

Références bibliographiques

Habert B., Nazarenko A., Salem A. (1997). *Les Linguistiques de Corpus*. - Colin.

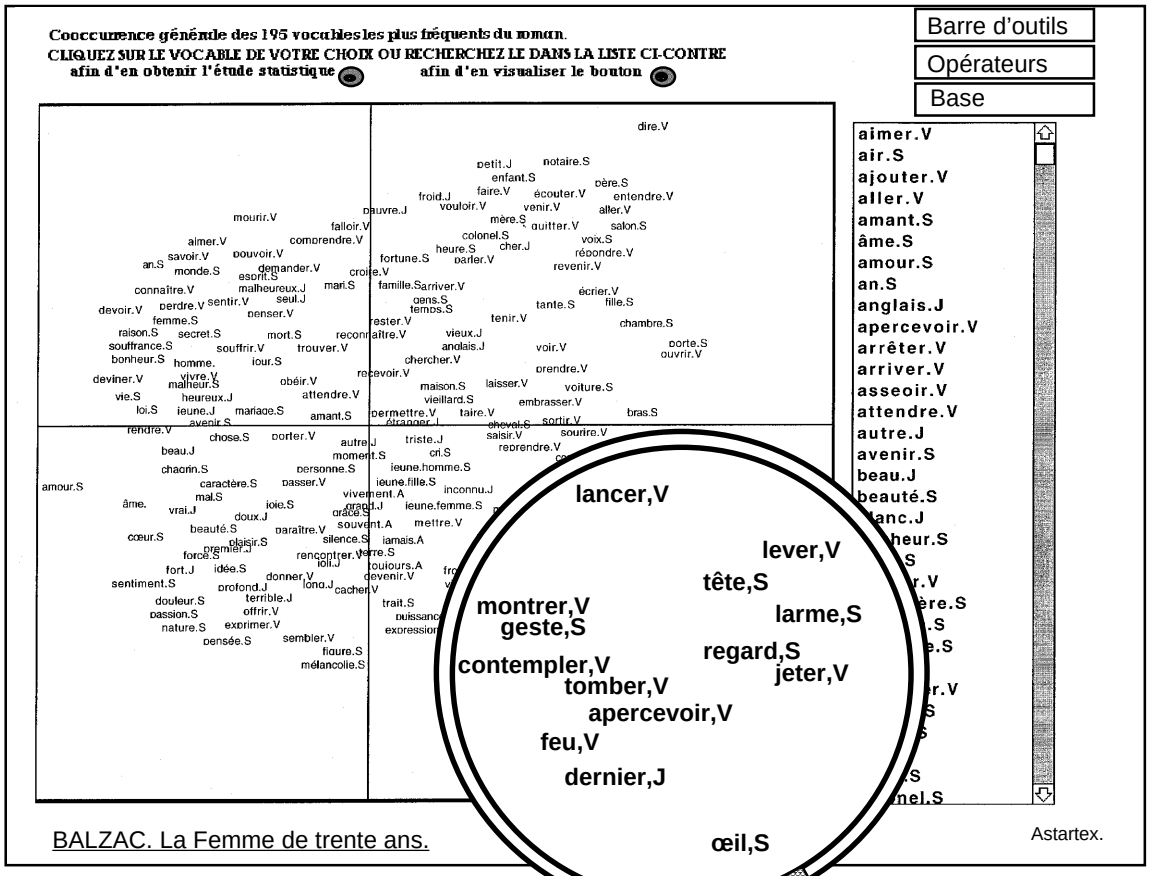
Lebart L., Salem A. (1994). *Statistique textuelle*. - Dunod.

Lelu A. (1998). Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de n-grammes. In *JADT 1998* - UNSA-CNRS, pages 391-400.

Viprey J.-M. (1997). *Dynamique du vocabulaire des* Fleurs du mal. - Champion.

Viprey J.-M. (1998). Une norme endogène pour le calcul stylistique du vocabulaire. In *JADT 1998* - UNSA-CNRS, pages 639-648

Viprey J.-M. (2000). *Théorie et méthodes d'analyse textuelle pour un hypertexte littéraire.* - Champion.


Cooccurrence générale des 195 vocables les plus fréquents du roman.
CLIQUEZ SUR LE VOCABLE DE VOTRE CHOIX OU RECHERCHEZ LE DANS LA LISTE CI-CONTRE
afin d'en obtenir l'étude statistique
afin d'en visualiser le bouton
Barre d'outils
Opérateurs
Base
aimer.V
air.S
ajouter.V
aller.V
amant.S
âme.S
amour.S
an.S
anglais.J
apercevoir.V
arrêter.V
arriver.V
asseoir.V
attendre.V
autre.J
avenir.S
beau.J
beauté.S
lancer,V
lever,V
tête,S
montrer,V
geste,S
larme,S
regard,S
contempler,V
jeter,V
tomber,V
apercevoir,V
feu,V
dernier,J
œil,S
BALZAC. La Femme de trente ans.
Astartex.