La linguistique de corpus à l'épreuve du numérique : textes, textures, documents

Rossana de Angelis

LA LINGUISTIQUE DE CORPUS À L'ÉPREUVE DU NUMÉRIQUE : TEXTES, TEXTURES, DOCUMENTS

Rossana DE ANGELIS
Céditec, Université Paris Est Créteil

Résumé

Les premières techniques d'exploitation des corpus ont été développées pendant les années 1970 au sein des recherches sur l'analyse lexicale. Au début des années 1990, les textes deviennent des objets d'analyse spécifiques afin d'identifier les modèles lexico-grammaticaux sous-jacents. Toutefois, à partir des années 2000, la technologie a changé le rapport aux textes. La possibilité de transformer un écrit dans un ensemble de données offre des nouvelles modalités d'exploitation, notamment en ce qui concerne le traitement automatique des données. En effet, l'impact des technologies numériques sur les méthodes d'exploitation des corpus provoque une remise en question du concept de *texte*, et par conséquent aussi une redéfinition du concept de *textualité*.

Mots clés

texte, textualité, corpus, matérialité numérique, herméneutique numérique

Abstract

The first techniques of corpus analysis were developed during the 1970s in the field of lexical analysis. In the early 1990s, the texts become some specific objects to explore and identify lexical and grammatical patterns underlying. However, from the 2000s, technology has changed the relationship to the texts. The possibility of transforming a written text in a data set offers new methods of exploring corpus, including as regards the automatic data processing. The impact of digital technologies on the data processing in corpus analysis provoques a new definition of the concept of *text*, and therefore also a redefinition of the concept of *textuality*.

Key words

texts, textuality, corpus, digital materiality, digital hermeneutics

0. INTRODUCTION

Les premières techniques d'exploitation des corpus ont été développées pendant les années 1970 au sein des recherches sur l'analyse lexicale[1]. Au début des années 1990, les textes deviennent ses objets spécifiques. À cette époque Sinclair (1991, p. xvii) analyse des corpus constitués de textes (*extended texts*) écrits en langue anglaise[2], afin

[1] Cf. Sinclair, Jones, Dayle 1970.

[2] "The *Describing English Language* series provides much-needed descriptions of modern English. Analysis of extended naturally-occurring texts, spoken and written, and, in particular, computer processing of texts have revealed quite unsuspected patterns of language." (Sinclair 1991, p. xvii).

d'identifier les modèles lexico-grammaticaux sous-jacents. Toutefois, comme le fait remarquer Meyer (2012, p. 24), la technologie a changé le rapport aux textes. La possibilité de transformer un écrit dans un ensemble de données permet une nouvelle modalité d'exploitation de celui-ci[3].

Le numérique propose une nouvelle vision du texte et de la textualité. « Un texte anti-naturel donc, dématérialisé – virtuel pourrait-on dire commodément –, dont les contours physiques tels que perçus depuis des siècles sont abolis, et la structure et le contenu – entendons, pour faire simple : la textualité – reconsidérés » (Mayaffre 2007a, p. 17). L'impact des technologies numériques sur les méthodes d'exploitation des corpus provoque une remise en question du concept de *texte.* Celle-ci se présente, par exemple, sous la forme du « *dépassement/complément de la linéarité* » (Mayaffre 2007a, p. 17), ce qui comporte par conséquent aussi une redéfinition du concept de *textualité*.

1. LA NOTION (LINGUISTIQUE) DE *TEXTE* AU MOMENT DE L'ÉMERGENCE DE LA LINGUISTIQUE DE CORPUS

Au sein de la *linguistique textuelle* telle qu'elle s'envisage au début des années 1970, le *texte* se présente en tant qu' « objet formel abstrait » (Slakta 1975, p. 30), s'opposant au *discours* considéré comme une « pratique sociale concrète » (Slakta 1975, p. 30). « Les phrases, les textes n'existent pas. Les énoncés, les discours sont des pratiques concrètes à analyser en relation aux autres pratiques sociales » (Slakta 1977, p. 20).

En s'inspirant de la réflexion de Slakta (1975, 1977), Adam (1977) propose la différentiation entre un « ordre du texte » et un « ordre du discours », s'achevant dans une distinction systématique entre les deux notions de *texte* et *discours* (Adam 1989, 1990, 1999). « Le texte est un *objet abstrait* résultant de la soustraction du contexte opérée sur l'*objet empirique* (discours) » (Adam 1989, p. 191). Cette différenciation lui permet de mettre en place la célèbre formule : « *Discours* = *Texte* + *Contexte* et *Texte* = *Discours* – *Contexte* » (Adam 1990, p. 23). Adam fera toutefois évoluer sa réflexion jusqu'à considérer cette formule comme « malheureuse » (Adam 2008, p. 30). « Il est, en effet, impossible de réduire les produits textuels des pratiques discursives des sujets – ce que nous nommerons *les textes* – à une telle organisation, objet abstrait d'une théorie *du texte* » (Adam 1999, p. 18). La formule proposée par Adam a toutefois été reprise par Jeandillou (1997, p. 108-109) au sein de l'*analyse textuelle*, approche qui répond justement à la nécessité de saisir les interactions entre ces deux ordres d'analyse.

Finalement, au sein du domaine hétérogène de la linguistique textuelle, on peut aborder le texte selon deux points de vue différents : l'un externe qui considère le texte comme « unité globale d'un acte d'énonciation » (Lundquist 1980, p. 1) ; l'autre interne qui considère le texte comme « l'enchaînement de structures syntaxiques particulières » (Lundquist 1980, p. 1).

En résumant, au sein de la linguistique textuelle, et au moment de l'apparition de la linguistique de corpus, le *texte* se présente de manière plutôt homogène en tant que « objet abstrait » (Van Dijk 1984). En particulier, il est envisagé comme le résultat

[3] "Technology, however, has greatly changed how we view the text: it is no longer an isolated entity existing only in printed form and accessible only through sometimes tedious manual analysis. Instead, when text is encoded in computer-readable form and becomes part of an electronic corpus, it can be annotated with linguistic information (e.g. all words can be assigned a part of speech act) and subjected to many different kind of linguistic analysis." (Meyer 2012, p. 24).

d'une « construction abstraite » (Van Dijk 1973, p. 179), une entité « construite par l'analyse » (Adam 2008, p. 44).

2. La matérialité des textes dans la linguistique de corpus

En travaillant sur les corpus, on peut distinguer trois lectures complémentaires à la lecture oculaire linéaire traditionnelle dès le moment où le texte est mis à l'épreuve du numérique : « [l]ecture quantitative, lecture paradigmatique (par le biais d'index notamment), lecture hypertextuelle (par le jeu des liens et des renvois), […] en complément de la lecture linéaire usuelle » (Mayaffre 2007a, p. 18). Cette pluralité de lectures est issue de la combinaison entre deux approches du texte linguistique :

– la linguistique de corpus mise à l'épreuve du numérique (ce qui suppose que le corpus soit constitué d'une pluralité de textes numériques) ;

– l'herméneutique matérielle (Szondi 1975) appliquée à l'analyse automatisée des textes linguistiques.

La rencontre entre ces deux approches en produit une nouvelle qui prend le nom de *philologie et/ou herméneutique numérique*. Autrement dit, et en reprenant les mots de Viprey (2005), l'objectif de cette nouvelle approche est de combiner la lecture linéaire à des lectures « *tabulaire* et *réticulaire* » tout en considérant la dimension matérielle des textes linguistiques comme une de leurs dimensions constitutives.

> De fait, les logiciels d'analyse de données textuelles, notamment ceux qui privilégient l'approche quantitative, commencent par faire exploser la linéarité du texte pour présenter leurs données en tableaux : tableaux alphabétiques, tableaux de fréquences, tableaux de distances, etc. Ces tableaux ne prétendent certes pas être le texte, mais ils sont une vision systématique et organisée – après l'explosion, le rangement – de la *matière textuelle* et deviennent les matrices sur lesquelles nos interprétations seront fondées.
>
> Plus subtilement, l'enjeu le plus complexe de l'analyse de données textuelles est de déceler les relations – relations autres que syntaxiques – que les items linguistiques entretiennent entre eux, non dans la phrase mais dans le texte en sa globalité. *Texte, textualité, texture* : l'objectif est de renouer avec l'étymologie même de ces mots et de démêler les trames et les entrelacs sous-jacents. Vision réticulaire donc des textes et des corpus qui met à jour les réseaux lexicaux pour (re)construire les thématiques, les isotopies ou isotropies récurrentes. (Mayaffre 2007a, p. 18-19, nous soulignons)

La linguistique de corpus appliquée dans un environnement numérique permet donc de formaliser cette vision du texte comme *texture*. L'analyse des corpus permet d'objectiver les co-textes et les con-textes des textes linguistiques dont ils se composent, « c'est-à-dire, comme des réseaux sémantiques auto-suffisants » (Mayaffre 2007a, p. 21). En adoptant cette perspective, « le corpus est la seule forme possible d'objectivation de l'intertexte » (Rastier 1998, p. 17) nécessaire à l'interprétation des textes constituants. « En un mot, les corpus numériques – par leur taille et leur organisation – doivent être élaborés et perçus comme des architextes sémantiques qui comprennent, en leur sein, les ressources textuelles nécessaires à leur compréhension/interprétation » (Mayaffre 2007a, p. 21).

En adoptant cette perspective, on voit donc que les concepts traditionnels de *texte* et *textualité* ne sont plus adéquats pour répondre aux exigences d'une linguistique de corpus dont les textes constituants offrent des nouvelles possibilités d'exploitation en raison de la *matière* toute particulière dont ils se constituent. De ce fait, « la textualité doit résolument être pensée comme la combinaison de parcours linéaires et réticulaires » (Adam 2006, p. 5).

La matérialité des textes linguistiques a un impact important sur les possibilité d'exploitation des textes par les méthodes de la linguistique de corpus.

> a computer-held corpus has to have the material in electronic form, either from print or obtained direct from a text-processing activity which uses computers (printing, word processing, electronique mail, etc.). There are three normal methods of text input at the present time:
> a. adaptation of material already in electronic from;
> b. conversion by optical scanning (machine reading);
> c. conversion by keyboarding. (Sinclair 1991, p. 14).

Néanmoins, pour pouvoir les exploiter à travers les outils de la linguistique de corpus, les textes manuscrits et/ou imprimés doivent être transformés dans une version électronique propre, c'est-à-dire sans aucun code supplémentaire.

> Once a text has been selected for study, the first decision is in what way it is to be re-created inside the computer. It may seem a simple enough process, to reproduce a text inside the machine, but in practice not all the features of a text are coded. Different features are picked out according to the need of the work. For most generally processing the text is kept to a very simple format – usually a single long string of letters, spaces, and punctuation marks. […] This is all, then, that the computer "knows" about text – a long succession of non described characters marked off in pages and lines (Sinclair 1991, p. 28).

Comme le rappelle Sinclair, ce ne sont pas tous les aspects matériels des textes qui sont codés par les machines. Et pourtant cette matérialité devient de plus en plus représentative. Par exemple, la différence entre les mots anglais *polish* et *Polish* - exemple rapporté par l'auteur (Sinclair 1991, p. 28) – peut être identifiée en enregistrant la différenciation matérielle entre les lettres minuscules et majuscules dans la transformation des textes en vue de la constitution du corpus[4].

L'attention réservée à la matérialité des textes reste encore aujourd'hui un des points faibles de la linguistique de corpus. En effet, des textes particuliers tels que ceux qui appartiennent au genre de la poésie visuelle ne peuvent pas être exploités sans trouver les moyens de coder la mise en page. La même difficulté a été relevée récemment à propos des textes multimodaux (Allwood 2008). Par exemple, la possibilité de constituer un corpus de bandes dessinées (les *manga*) suppose un travail préalable de codification de l'espace graphique, en envisageant ainsi la possibilité de constituer un corpus orienté (*text-oriented corpus*, Unser-Schutz 2011) par la nature même des textes considérés selon leur propre complexité qui dépend notamment de la dimension visuelle de la mise en page[5].

Le problème que représente le traitement des éléments non (strictement) linguistiques en tant qu'éléments linguistiquement significatifs reste ouvert. Alors que les techniques

---

[4] "Using the sample notion of word-form, we can now represent a text as a succession of word-forms. The word-forms can be counted, so that the length of the text, measured in word-forms, can be calculated. Next, the word-forms can be compared with each other, and it will be found that there are many repetitions of the same word-form. So another count can be made of the number of different word-forms, which is called the *vocabulary* of the text." (Sinclair 1991, p. 29)

[5] "Ultimately this is a text-oriented corpus, insofar as it is an attempt to deal with how to think about and link together non-linear text which is by nature a part of a larger visual structure." (Unser-Schutz 2011, p. 214)

de codage des éléments purement linguistiques se sont affinées, les autres se développent très lentement (Bateman, Delin, Henschel 2002), ce qui pénalise évidemment l'analyse des documents multimodaux[6].

Toutefois, paradoxalement, les problèmes posés par la matérialité des textes linguistiques deviennent évidents dès qu'on envisage l'exploitation d'un corpus constitué de textes numériques à l'aide des dispositifs d'analyse automatique des textes linguistiques.

Caro Dambreville (2007) qualifie le document numérique d'immatériel et de « virtuel ». En effet, sur le support d'enregistrement les signes ne sont pas directement lisibles, ce qui rend leur existence « virtuelle » en l'absence de dispositif de décodage (Balpe 1990). « Le document papier, lui, porte l'information de manière indissociable, support de présentation et information ne faisant qu'un pour le lecteur, l'information étant en quelque sorte "incrustée dans le support" » (Caro Dambreville 2007, p. 46). La différence entre les textes traditionnels et les textes numériques consiste premièrement – mais pas seulement – dans une relation différente à leur support : le texte imprimé ne met jamais en suspens la matérialité de son support car – comme le disait Genette (1987) – le support *présente* le texte dans le double sens de le rendre présent et le montrer au lecteur ; en revanche, le texte numérique met en suspens la matérialité du support dès qu'on soustrait le texte au dispositif, ce qui rend le texte numérique « virtuel ».

3. La matérialité des textes numériques

Selon les approches traditionnelles, le *texte* est ce qui assure à des éléments sémio-linguistiques « une existence concrète, matérielle » (Adam, Goldstein 1976, p. 195). Toutefois, les *textes numériques* sont normalement considérés comme des textes dématérialisés. Ce qui manque est donc « une prise en compte de la matérialité propre du numérique » (Doueihi 2011, p. 301). En effet, les interfaces qui rendent ces textes visibles et lisibles supposent d'adapter non seulement nos pratiques au changement de support, mais aussi nos habitudes perceptives en tant que lecteurs (Dacos, Mounier 2010 ; Sinatra, Vitali-Rosati 2014).

Toutefois, la dimension matérielle des textes numériques peut être prise en charge par une *herméneutique numérique*. En fait, le changement du support – comme du papier au numérique, par exemple – demande le changement des pratiques d'exploitation et d'interprétation des textes. Une approche herméneutique, et notamment une *herméneutique matérielle* inspirée de l'œuvre de Szondi (1975), permet d'envisager la dimension matérielle des textes pendant leur analyse.

> L'herméneutique numérique est une herméneutique matérielle ; pas seulement par conviction mais par nécessité. Ou plutôt : avec le numérique l'évidence devient nécessité. […] la machine en effet ne saurait embrasser le texte autrement que par sa matière. Sauf à renverser le procès de la démarche et s'illusionner sur les possibilités de l'intelligence artificielle, l'ordinateur ne peut donner accès au sens d'un texte sans appréhender sa

[6] "For multimodal documents, however, such models and techniques are still largely lacking. A substantial set of problems is raised by the fact that the object of study is not linear, either temporally or in terms of the principles for its consumption; moreover, its multichannel nature makes it difficult to reconcile and peg together the methods of recording, transcription, analysis and annotation that have been developed separately for each mode." (Bateman, Dlin & Henschel 2002, p. 3)

> lettre ; il ne saurait aborder son esprit sans traiter (« saisir », « implémenter », « digitaliser », « numériser ») sa matière. (Mayaffre 2007a, p. 22).

Selon cette perspective, « la philologie et/ou herméneutique numérique révolutionnent non seulement notre rapport aux textes et à la textualité, mais aussi nos pratiques heuristiques quotidiennes, mais encore, tout simplement, nos connaissances et notre appréhension de la culture (textuelle) humaine » (Mayaffre 2007a, p. 16). En effet, les nouvelles modalités d'exploitation du texte à travers les différentes déclinaisons du Traitement Automatique des Langues (TAL) permettent de relever la connexion existante entre l'identification matérielle des éléments textuels – tels que, par exemple, les (co)occurrences des syntagmes sélectionnés – et l'identification des parcours interprétatifs (Rastier 2009). Ainsi les extraits d'un corpus sélectionnés par le biais de procédures automatiques d'extraction des données textuelles constituent en quelque une sorte de « nouveau » texte.

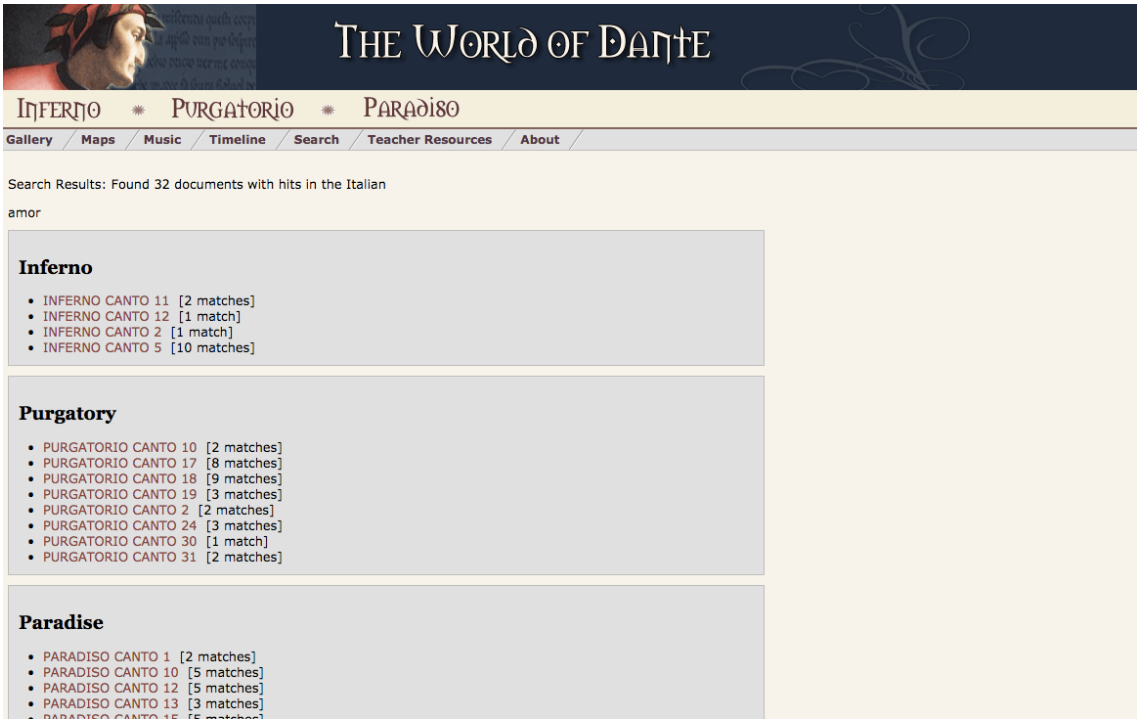

Image 1. Capture d'écran des résultats de la recherche du terme "amor"

Voici un exemple. Le site *The World of Dante*[7] est un outil de recherche multi médias qui facilite l'étude de la *Divina Commedia* écrite par Dante Alighieri, chef d'œuvre de la littérature internationale dont l'analyse est particulièrement complexe. Ceci met à disposition une transcription du texte italien, sa traduction anglaise, des diagrammes explicatifs, une ligne du temps, des galeries d'images, etc. Le texte italien encodé

[7] http://www.worldofdante.org/

permet « des recherches et des analyses structurées » – comme on peut lire sur le site – et, pour cette même raison, une approche dynamique, en pouvant ainsi intégrer ces différentes composantes pendant l'analyse de l'œuvre. Le texte de la *Divina Commedia* est exploitable à tout niveau : grâce aux outils de recherche disponibles on peut identifier les éléments textuels comme, par exemple, un mot, un personnage, un lieu, etc. en cherchant ses occurrences dans le texte, en analysant le co-texte et le contexte où ils apparaissent, en faisant le lien entre les éléments textuels saisis par les outils d'exploitation automatique des données textuelles avec les éléments extra-textuels auxquels ils sont en relation dans le site (par exemple, les images représentant les lieux et/ou les personnages présents dans le texte, les musiques de l'époque, et les informations les plus variées).

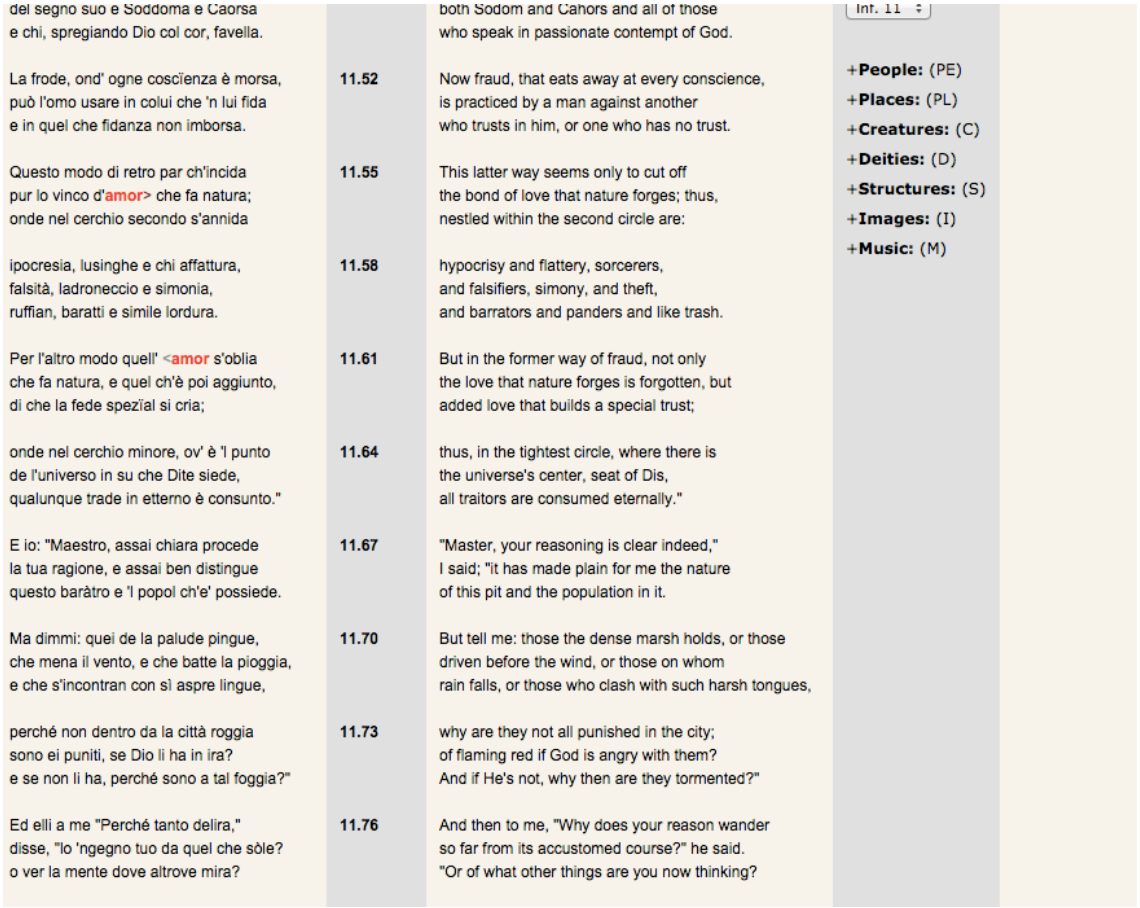

Image 2. Capture d'écran des quelques occurrences du terme "amor" retrouvées suite à la recherche lancée

En utilisant ces outils de recherche, on peut découvrir à la fois les (co)occurrences des éléments textuels choisis et les liens intra- et inter- textuels qu'ils entretiennent à l'intérieur et à l'extérieur du texte lui-même avec d'autres éléments (non) textuels. Autrement dit, on peut constater que *les liens intra- et inter- textuel se matérialisent* en

se visualisant sur l'écran grâce à des outils d'exploitation qui cherchent les occurrences de ces éléments dans la chair numérique du texte.

Une autre conception du texte se rend donc visible à nos yeux : il ne s'agit pas du *texte* selon la conception propre à la culture des textes imprimés, mais de la *textualité* conçue comme l'ensemble des liens intra- et inter-textuels qui se visualisent à l'écran[8]. Et finalement, la matérialisation de la *textualité* rétroagit sur la conception même du *texte*.

Ce passage suppose donc un changement de paradigme *:* en passant de la culture des textes imprimés à celle des textes numériques, ce qu'on identifie comme *texte* se propose comme une véritable *texture* grâce à la matérialisation du réseau d'éléments intra- et inter- textuels par des nouvelles pratiques d'exploitation du corpus. Ces éléments acquièrent une réalité (numérique) sous nos yeux, ils se matérialisent en changeant la conception même du *texte*.

4. UNE APPROCHE HERMÉNEUTIQUE DE L'EXPLOITATION DES CORPUS NUMÉRIQUES

L'*herméneutique numérique* se présente comme une approche issue de l'analyse de corpus constitués de textes numériques. « Chevillée donc à la matière textuelle » (Mayaffre 2007a, p. 19), cette approche défend la possibilité de proposer une interprétation objective des textes à travers l'identification de parcours interprétatifs (Rastier [1987] 2009) en analysant les textes à l'aide d'outils d'exploitation automatique des données textuelles qui permettent de repérer et de tracer les liens entre les différentes occurrences des éléments textuels identifiés.

> C'est en ce sens qu'elle peut se revendiquer de Peter Szondi et de son herméneutique critique ; c'est en ce sens que l'on parle d'une herméneutique philologique. Les parcours interprétatifs sont toujours sujets à caution, mais la trajectoire de ceux de la philologie et/ou herméneutique numérique a l'avantage d'être solidement inscrite dans la bonne direction grâce à son décisif et premier mouvement : par la prise en compte nécessaire, systématique et exhaustive, des matériaux linguistiques (lettres et syllabes, formes graphiques et lemmes, codes grammaticaux et enchaînements syntaxiques, segments répétés, expressions, cooccurrents, collocations micro-distributionnelles, réseaux lexicaux, concordances phrastiques, contextes paragraphiques, etc.) des textes. (Mayaffre 2007a, p. 19)

Ainsi, « le numérique en multipliant les mises en forme des textes propose une autre vision du texte » (Mayaffre 2007a, p. 17). L'aspect le plus novateur de ce changement est représenté par « *le dépassement/complément de la linéarité* » (Mayaffre 2007a, p. 17). La dé-linéarisation est une première conséquence des différents modes d'exploitation des textes à l'ère du numérique. Le traitement automatique des données s'applique notamment à l'analyse de corpus constitués de textes dont l'analyse suppose l'identification de relations significatives *intra-*, *inter-* et *archi-* textuelles (Genette 1982, 1987). L'analyse assistée par ordinateur suppose ainsi l'identification des relations sémantiques entre les différents éléments textuels, à l'intérieur mais aussi à l'extérieur du corpus lui-même. Celle-ci peut se réaliser selon des procédures différentes : *lexicométrie*, *textométrie*, *logométrie*. Ces approches supposent une

[8] « Le texte graphique n'est guère plus qu'un texte lemmatisé, un texte objectif ou naturel. […] Grâce à la performance des lemmatiseurs/étiqueteurs, et malgré leurs erreurs résiduelles, la surface lemmatisée ou grammaticalisée du texte n'est, aujourd'hui, guère moins contestable ou arbitraire que celle d'un texte brut » (Mayaffre 2007a, p. 8, n. 1).

conception fermée du corpus, envisagé comme un hypertexte, et une conception ouverte du *texte* : conçu au sein d'un corpus, un texte n'est jamais isolé, mais sa propre identité s'affirme en relation aux autres textes compris dans le corpus constitué.

La *lexicométrie*, par exemple, est une des pratiques d'analyse dont se sert notamment la linguistique de corpus : « organiser le retour au texte pour en permettre la lecture et favoriser l'acte final interprétatif est une de ses tâches fondamentales » (Mayaffre 2008, p. 92). En supposant les modalités de fonctionnement d'une statistique lexicale[9], le texte est considéré comme le seul objet linguistique empirique qui peut être soumis à l'analyse (Rastier 2009). L'objectif est de montrer que le sens dépend des parcours (inter)textuels qui se dessinent pendant l'analyse, en supposant ainsi que les textes se trouvent en amont et en aval de l'analyse linguistique.

Selon cette approche, « les corpus sont des objets construits […] qui informent leurs composants » (Mayaffre 2008, p. 94, n. 5), c'est-à-dire les textes. L'analyse des textes se fait donc à travers la (re)construction de *parcours interprétatifs* (Rastier 2009) qui sont à la fois des parcours *textuels* (identifiés à l'intérieur de chaque texte constituant le corpus), *co-textuels* (identifiés dans l'environnement linguistique) et *con-textuels* (identifiés grâce aux « co-occurrences » qui représentent des formes minimales de contexte supposant à la fois des occurrences dans le texte – conçu comme unité globale au palier inférieur de l'analyse – et dans le corpus – conçu comme unité globale au palier supérieur de l'analyse –). Ainsi, « les travaux les plus novateurs d'ADT [analyse de données textuelles] visent à compléter l'approche statistique paradigmatique ou non-séquentielle originelle de la lexicométrie par un traitement plus global de la surface des textes et des corpus, à même de rendre compte de leur organisation spatiale, linéaire ou continue : […] leur organisation *topographique* ou *topologique* » (Mayaffre 2007b, p. 3).

De son côté, la *textométrie* a développé des nouvelles techniques pour analyser des grands corpus de textes. En reprenant les modalités d'analyse de la *lexicométrie*, elle propose en particulier une analyse de corpus statistiquement fondée. Comme l'explique Pincemin (2011, 2012), la textométrie développe des outils pour analyser notamment les corpus numérisés. Cette approche, développée dans le cadre de la sémantique interprétative (Rastier 2009), montre comment la cooccurrence des éléments textuels mesurée par la textométrie (considérée comme une approche quantitative) puisse se révéler efficace pour la caractérisation des textes et des genres textuels. Le projet nommé Textométrie[10], par exemple, concerne le développement de logiciels *open-source* pour mettre en place une plateforme modulaire appelée TXM[11].

Une autre approche de l'analyse des corpus selon les modalités d'extraction automatique des données textuelles est la *logométrie*, fondée sur la combinaison de deux modalités différentes par lesquelles pouvoir interpréter les résultats de l'analyse :

---

[9] « La volonté de remettre les formes dans leur contexte se caractérise, en lexicométrie, par deux types de comportement de l'analyste et deux modes de fonctionnalités classiques des logiciels : le *retour au texte*, simple mais essentiel, et le développement d'une *statistique contextualisante*, syntagmatique ou co-occurrencielle. » (Mayaffre 2008, p. 91).

[10] Cf. Heiden, S., Magué, J-P., Pincemin, B., 2010.

[11] « Il s'agit à la fois d'une opération patrimoniale au rayonnement international et du lancement d'une nouvelle génération de recherche textométrique, en synergie avec les technologies de corpus actuelles (Unicode, XML, TEI, outils de TAL, CQP, R) ». http://textometrie.ens-lyon.fr/?lang=fr

> En lisant le corpus différemment (lecture discontinue – mais systématique – complémentaire à la lecture continue – mais aléatoire, selon la concentration –, lecture paradigmatique complémentaire à la lecture syntagmatique, lecture tabulaire et réticulaire complémentaire à la lecture linéaire, lecture quantitative complémentaire à la lecture qualitative, lecture hypertextuelle complémentaire à la lecture textuelle), l'ordinateur interroge différemment. (Mayaffre 2012, p. 28)

En utilisant les différentes modalités de l'analyse statistique des données textuelles, la *logométrie* propose donc une approche « objective » dans l'interprétation des textes constituants le corpus en contrôlant les parcours de lecture. Selon cette perspective, en fait, « le corpus est la seule forme possible d'objectivation de l'intertexte » (Rastier 1998, p. 17) immédiatement nécessaire à l'interprétation des textes qui le constituent (Mayaffre 2007a, p. 21).

5. TEXTES, TEXTURES, DOCUMENTS

La culture numérique renvoie une autre vision du *texte* : ses limites disparaissent, la textualité vient au premier plan, et il nous est requis de changer nos habitudes perceptives et interprétatives pour pouvoir l'analyser. Néanmoins, comme on vient de le voir, ce qui saisit tout de suite notre attention est le processus de dé-linéarisation qui se met en place à l'ère du numérique.

« Suite » (Rastier 2001, p. 21) d'éléments linguistiques « progressant vers une fin » (Détrie, Siblot, Vérine, 2001, p. 349), « unité communicative » (Weinrich 1976) qui fait preuve de « cohérence et cohésion » (De Beaugrande, Dressler 1981), « plan » du langage (Adam 1999, p. 5). Alors que la plupart des définitions linguistiques du texte utilisées normalement insiste sur sa dimension linéaire, les nouvelles approches des textes numériques remettent en question justement cette linéarité comme étant constitutive. Au sein de la culture numérique, comme le montre notamment la linguistique de corpus à l'aide des dispositifs d'analyse automatique des données, « la textualité doit résolument être pensée comme la combinaison de parcours linéaires et réticulaires » (Adam 2006, p. 5). Dans la culture numérique le *texte* devient de plus en plus une *texture.*

Au début des années 1990, la linguistique de corpus marquait une différence très nette entre les concepts de *textes* et de *documents*[12], distinction reprise encore aujourd'hui[13]. Les pratiques de numérisation remettent cependant en question aussi la définition de *document*[14], ce qui amène naturellement à réviser la relation entre ces concepts.

---

[12] "Not many features of a book-length text are diffused evenly throughout, and a corpus made up of whole documents is open to a wider range of linguistic studies than a collection of short samples." (Sinclair 1991, p. 19)

[13] « La structure typographique de la mise en forme relève d'un niveau de l'expression textuelle ; en revanche, la mise en page relève de la norme du document – même si elle n'est pas sans effet sur l'appréhension du texte. Ainsi les paginations comme les titres courants font-ils partie du document, mais non du texte. À la philologie numérique répond ainsi une *diplomatique* numérique, qui ne traite pas du texte, mais seulement de caractères spécifiques au document qui le véhicule » (Rastier 2011, p. 68).

[14] À titre d'exemple, « le document numérique se dépouille des qualités du document unique de l'archiviste : authentifiable, doué par sa continuité matérielle d'une intégrité (même quand il est fragmentaire), non reproductible, faisant autorité. L'affichage par pixel détruit toute continuité matérielle qui empêchait les falsifications. Alors qu'une critique initiale suffisait à établir ce

D'une façon générale, les *documents* peuvent être considérés comme des objets linguistiques produits par des *pratiques* particulières. Comme on vient de le voir, l'*herméneutique matérielle* (Szondi 1975, Molinié 2005, Mahyew 2007) exploite le texte en l'envisageant comme un objet linguistique produit par une pratique. « En réunifiant l'herméneutique et la philologie, l'*herméneutique matérielle* place la problématique de l'interprétation au centre des sciences du langage » (Rastier 2001, p. 99).

Toutefois, l'*herméneutique numérique* est une *herméneutique matérielle* car, comme on vient de le voir, les logiciels d'analyse automatique des données textuelles donnent accès aux textes constituant un corpus en traitant justement la matière textuelle : mots, caractères, espaces, etc. par des processus de lemmatisation nécessaires à l'analyse.

> L'ordinateur décompose ses objets en plus petites unités sémiotiques. Et un corpus est pour lui d'abord constitué de lettres concaténées, de blanc et de ponctuation, d'octets et de bits. Si ces unités sont ensuite combinées, reliées, contextualisées (voir interprétées comme dans le cadre de la lemmatisation), la description comme l'interrogation numérique du texte s'appuieront sur ces signaux informatiques premiers et minimaux. (Mayaffre 2007a, p. 24).

L'attention réservée à la matérialité des textes au sein des corpus numériques change nos pratiques analytiques. Par exemple, au plan graphique, « [a]lors que la *ponctuation* n'est pas considérée comme sémantique et qu'elle est tout simplement absente des grammaires formelles, l'étude en corpus permet de souligner les corrélations entre contenus lexicaux et ponctèmes » (Rastier 2005, p. 38).

Finalement, plutôt qu'être *dé-matérialisé*, le texte numérique soumis au traitement automatique devient *sur-matérialisé :* dès le moment où le traitement dépend de la composition matérielle du texte, et de sa présence matérielle dans le corpus, sa matérialité acquiert une importance capitale. Le statut du *texte* se rapproche de celui du *document*… où sont donc les limites ?

La numérisation au sein de la linguistique de corpus change non seulement la notion de *texte* en faveur de celle de *texture*, mais aussi la relation entre les notions de *texte* et de *document*. En effet, ce dernier suppose normalement de pouvoir distinguer deux aspects : l'expression du texte et la configuration du support. Les textes numériques remettent en question cette distinction. Et la numérisation rapproche ainsi le *texte* du *document*.

En réfléchissant sur les nouvelles approches au sein de linguistique de corpus mise à l'épreuve du numérique, on remet en question finalement certains concepts fondamentaux des sciences du langage. Une réflexion sur la relation entre pratiques et objets s'ouvre ainsi au sein de la linguistique de corpus grâce à la prise en compte de la matérialité numérique.

---

document, il faut à présent une critique indéfinie pour maintenir une fiabilité. L'établissement des significations doit souvent passer par une succession de versions, dont chacune est le support et le résultat d'une opération de lecture. » (Rastier 2001, p. 19)

BIBLIOGRAPHIE

ADAM, Jean-Michel, 1977. « Ordre du texte, ordre du discours », *Pratiques*, 13, 103-111.

ADAM, Jean-Michel, 1989. « Pour une pragmatique linguistique et textuelle », *in* Reichler, Claude (éd.), *L'interprétation des textes*, Paris, Minuit, 183-222.

ADAM, Jean-Michel, 1990. *Éléments de linguistique textuelle*, Liège, Pierre Mardaga.

ADAM, Jean-Michel, 1999. *Linguistique textuelle. Des genres de discours aux textes*, Paris, Nathan.

ADAM, Jean-Michel, 2006. *Autour du concept de texte. Pour un dialogue des disciplines de l'analyse de données textuelles*, *JADT 2006*, disponible sur *Lexicométrica* [en ligne]. http://www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_JMA.pdf. Consulté le 20 mai 2016.

ADAM, Jean-Michel, 2008. *La linguistique textuelle. Introduction à l'analyse textuelle des discours*, Paris, Armand Colin [1re éd. *Sciences du texte et analyse du discours*, Genève, Slatkine Érudition, 2005].

ADAM, Jean-Michel & GOLDSTEIN, Jean-Pierre, 1976. « Vers une grammaire de texte », *Linguistique et discours littéraire : théorie et pratique des textes*, Paris, Larousse, 185-225.

ALLWOOD, Jens, 2008. « Multimodal corpora », *in* A. Lüdling et M. Kytö (ed.), *Corpus linguistics : an international handbook*, Berlin, Mouton de Gruyter, 207-225.

BALPE, Jean-Pierre, 1990. *Hyperdocuments hypertextes hypérmédias*, Paris, Eyrolles.

BATEMAN, John ; DELIN, Judith ; HENSCHEL, Renate, 2002. "Multimodality and empiricism: methodological issues in the study of multimodal meaning-making", *GeM report* 2002/1, 1-35.

CARO DAMBREVILLE, Stéphane, 2007. *L'écriture des documents numériques : approche ergonomique*, Paris, Lavoisier.

DACOS, Marin & MOUNIER, Pierre, 2010. *L'édition éléctronique*, Paris, La Découverte.

DALBERA, Jean-Philippe, 2002. « Le corpus entre données, analyse et théorie », *Corpus* [En ligne], 1/2002. http://corpus.revues.org/10. Consulté le 20 mai 2016.

DE BEAUGRANDE, Robert-Alain & DRESSLER, Wolfgang Ulrich, 1981. *Einführung in die Textlinguistik*, Tübingen, Niemeyer.

DÉTRIE, Catherine ; SIBLOT, Paul ; VÉRINE, Bertrand (éd.), 2001. *Termes et concepts pour l'analyse du discours. Une approche praxématique*, Paris, Champion.

DOUEIHI, Milad, 2011. *La grande conversion numérique*, Seuil, Paris [1re éd. 2008].

GENETTE, Gérard, 1982. *Palimpsestes. La littérature au second degré*, Paris, Seuil.

GENETTE, Gérard, 1987. *Seuils*, Paris, Seuil.

HEIDEN, Serge ; MAGUÉ, Jean-Philippe ; PINCEMIN, Bénédicte, 2010. « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », Bolasco, Sergio (éd.), *Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, vol. 2, p. 1021-1032. *Edizioni Universitarie di Lettere Economia Diritto*, Roma, Italy.

Disponible sur : halshs.archives-ouvertes.fr/halshs-00549779/fr/ Consulté le 20 mai 2016.

JEANDILLOU, Jean-François, 1997. *L'analyse textuelle*, Paris, Armand Colin.

LUNDQUIST, Lita, 1980. *La Cohérence textuelle : syntaxe, sémantique, pragmatique*, Copenhague, Nyt Nordisk Forlag Arnold Busck.

MAYAFFRE, Damon, 2002a. « L'Herméneutique numérique », *L'Astrolabe. Recherche littéraire et Informatique*.
Disponible sur : http://www.uottawa.ca/academic/arts/astrolabe/ Consulté le 20 mai 2016.

MAYAFFRE, Damon, 2002b. « Les corpus *réflexifs :* entre architextualité et hypertextualité », *Corpus* [En ligne], 1/2002.
http://corpus.revues.org/11. Consulté le 20 mai 2016.

MAYAFFRE, Damon, 2010. « « Corpus et web-corpus. Réflexion sur la corporalité numérique », *Cahiers de praxématique* [En ligne], 54-55 | 2010, document 13, mis en ligne le 01 janvier 2013.

http://praxematique.revues.org/1170 Consulté le 28 mars 2016.

MAYAFFRE, Damon, 2007a. « Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques », *in* Rastier, F. & Ballabriga, M. (éd.), *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation*, Toulouse, PUT, 15-26.

MAYAFFRE, Damon, 2007b. « L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan : Retour sur les travaux actuels de topographie/topologie textuelle. *Lexicometrica*, André Salem, Serge Fleury, 2007, p.1-12.

MAYAFFRE, Damon, 2007c. « L'analyse de données textuelles aujourd'hui : du corpus comme une urne, au corpus comme un plan. Bilan sur les travaux actuels de topographie/topologie textuelle », *Lexicométrica : Topographie et topologie textuelles* [en ligne].
http://lexicometrica.univ-paris3.fr/numspeciaux/special9/mayaffre.pdf. Consulté le 20 mai 2016.

MAYAFFRE, Damon, 2008. « L'entrelacement lexical des textes, co-occurrences et lexicométrie », *Texte et corpus*, n. 3 / août 2008, Actes des Journées de la linguistique de Corpus 2007, p. 91-102.
http://web.univ-ubs.fr/corpus/jlc5/ACTES/ACTES_JLC07_mayaffre.pdf Consulté le 20 mai 2016.

MAYAFFRE, Damon, 2012. *Nicolas Sarkozy. Mesure et démesure du discours (2007-2012)*, Paris, Presses de la Fondation National des Sciences Politiques.

MAYHEW, Robert J., 2007. "Materialistic hermeneutics, textuality and the history of geography: print spaces in British geography, c. 1500-1900", *Journal of Historical Geography*, 33, 466-488.

MAZIÈRE, Francine, 2005. *L'analyse du discours: histoire et pratiques*, Paris, PUF (Que sais-je ?).

MEYER, Charles F., 2012. "Textual analysis: from philology to corpus linguistics", Kytö Merja (éd.), *English Corpus Linguistics: Crossing Paths*, Amsterdam - New York, Rodopi, 23-42.

MOLINIÉ, Georges, 2005. *Hermès mutilé. Vers une herméneutique matérielle. Essai de philosophie du langage*, Paris, Honoré Champion.

PINCEMIN, Bénédicte, 2011. « Sémantique interprétative et textométrie – Version abrégée », *Corpus*, 10. https://corpus.revues.org/2121. Consulté le 20 mai 2016.

PINCEMIN, Bénédicte, 2012. « Sémantique interprétative et textométrie », *Texto !*, vol. XVII, 3. http://www.revue-texto.net/docannexe/file/3049/pincemin_texto11.pdf. Consulté le 20 mai 2016.

RASTIER, François, 1998. « Herméneutique matérielle et artéfacture. Échange entre François Rastier et Bruno Bachimont sur sa thèse *Herméneutique matérielle et artéfacture : Des machines qui pensent aux machines qui donnent à penser*», *Texto!*, décembre 1998. http://www.revue-texto.net/Dialogues/Rastier-Bachimont.html. Consulté le 20 mai 2016.

RASTIER, François, 2001. *Arts et sciences du texte*, Paris, Presses Universitaires de France.

RASTIER, François, 2005. « Enjeux épistémologiques de la linguistique de corpus », *in* Williams, J. (éd.), *La linguistique de corpus*, Rennes, PUR, 31-45. Disponible sur *Texto !*, juin 2004 : http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html. Consulté le 20 mai 2016.

RASTIER, François & BALLABRIGA, Michel (dir.), 2007. *Corpus en Lettres et Sciences sociales – Des documents numériques à l'interprétation*, Toulouse, Presses de l'Université de Toulouse Le Mirail.

RASTIER, François, 2009. *Sémantique interprétative*, Paris, PUF [1re éd. 1987].

RASTIER, François, 2011. *La mesure et le grain : sémantique de corpus*, Paris, Honoré Champion.

SINATRA, Michaël E. et VITALI-ROSATI, Marcello (éd.), 2014. *Pratiques de l'édition numérique*, Presses Universitaires de Montréal.

SINCLAIR, John, 1991. *Corpus, Concordance, Collocation*, Oxford, OUP.

SINCLAIR, John ; JONES, Susan ; DAYLE, Robert, 1970. *English Lexical Studies*, OSTI Report.

SLAKTA, Denis, 1975. « L'ordre du texte », *Études de linguistique appliquée*, n. 19, 30-42.

SLAKTA, Denis, 1977. *Introduction à la grammaire de texte. Actes de la session de linguistique de Bourg-Saint-Maurice, publications du conseil scientifique de la Sorbonne Nouvelle-Paris III, 4-8 septembre 1977*, Paris, Sorbonne Nouvelle, III, 7-63.

SZONDI, Peter, 1975. *Einführung in die literarische Hermeneutik*, Jean Bollack, Helen Stierlin (ed.), Frankfurt-am-Main, Suhrkamp. [trad. fr. *Introduction à l'Herméneutique Littéraire. De Chladenius à Schleirmacher*, Paris, Cerf, 1989].

UNSER-SCHUTZ, Giancarla, 2011. "Developing a text-based corpus of the language of Japanese comics (*manga*)", *in* Newman, John, Baayen, Harald, Rice, Sally (ed.), *Corpus-based Studies in Language Use, Language Learning and Language Documentation*, Rodopi B. V., Amsterdam - New York, NY, 213-238.

VAN DIJK, Teun Adrianus, 1973. « Grammaires textuelles et structures narratives », *in* Chabrol, Claude (éd.), *Sémiotique narrative et textuelle*, Paris, Larousse, 177-207.

VAN DIJK, Teun Adrianus, 1984. « Texte », *in* Beaumarchais, Jean-Pierre ; Couty, Daniel ; Rey, Alain (éd.), *Dictionnaire des littératures de langue française*, Paris, Bordas, 2281-2289.

VIPREY, Jean-Marie, 2005. « Philologie numérique et herméneutique intégrative », Adam, Jean-Michel et Heidmann, Ute (éd.), *Sciences du texte et analyse de discours*, Genève, Slatkine, 51-68.

WEINRICH, Harald, 1976, *Sprache in texten*, Stuttgart, Enrst Klett.

WILLIAMS, Geoffrey (éd.), 2005. *La linguistique de corpus*, Rennes, PUR.