CULTURONOMICS: JUSTE UNE QUESTION DE CORPUS?

LE 11 JANVIER 2011 ***OLIVIER ERTZSCHEID***

Nos sociétés de données nourrissent des monstres calculatoires et industriels qui, dans certains domaines, sont en passe d'être les seuls capables de circonscrire des corpus qui relèvent, pourtant, du bien commun. Génomique aujourd'hui, linguistique demain et la culture après-demain?

À quoi sert de numériser des millions d'ouvrages depuis 2005 ? À **ça** (« Quantitative Analysis of Culture Using Millions of Digitized Books », article publié dans la revue scientifique *Science*). Disposer de 4% de tous les livres publiés depuis 2 siècles. 7 langues. 2 milliards de mots. 5,2 millions de livres numérisés "inside" (voir **l'article du NYTimes**).

Ici (Google), **le plus grand corpus linguistique de tous les temps**.
Ailleurs (Facebook), **le plus grand "corp(u)s social" numérique**.
Deux corpus. Mais qu'est-ce qu'un corpus ?

"*Ensemble de données exploitables dans une expérience d'analyse ou de recherche automatique d'informations.*" (Source : **Trésor de la langue française**)
"*Ensemble de textes établi selon un principe de documentation exhaustive, un critère thématique ou exemplaire*" (Source : **Trésor de la langue française**)
Dans le **domaine du droit**, le corpus : "*C'est l'élément matériel de la possession, le pourvoir de fiat exercé sur une chose. (Animus).*"

Du premier corpus, celui de Google, on ne pourra que se réjouir, pour ce qu'il représente de potentialités ouvertes dans l'aventure linguistique comme compréhension du monde. Et l'on mettra du temps à en épuiser les possibles. Mais nul doute qu'il contribuera aussi à alimenter tous les fantasmes, celui, notamment, d'une "intelligence artificielle" dévoyée, apprenant à penser en déchiffrant ce que le plus grand corpus du monde révèle des pensées de ce même monde. Les ingénieurs ont même inventé un mot pour cela : "culturonomics". Culture et génomique. Enthousiasmant. Pour l'instant. Et pour les linguistes.

Du second corpus, celui de **Facebook**, on ne peut que continuer à raisonnablement s'alarmer. Surtout lorsque les techniques de traitement dudit corpus prennent **cette orientation**, rendant plus que jamais nécessaire la mise en œuvre d'un littéral **Habeas Corpus** numérique.

Dans l'histoire des sciences, les scientifiques de tous les domaines, de toutes les époques,

de toutes les disciplines, se sont en permanence efforcés de prendre l'ascendant sur leur différents corpus ; pour pouvoir être exploitable, le corpus doit pouvoir être circonscrit par ceux qui prétendent en faire l'analyse.

> ***Il n'y a rien que l'homme soit capable de vraiment dominer : tout est tout de suite trop grand ou trop petit pour lui, trop mélangé ou composé de couches successives qui dissimulent au regard ce qu'il voudrait observer. Si ! Pourtant, une chose et une seule se domine du regard : c'est une feuille de papier étalée sur une table ou punaisée sur un mur. L'histoire des sciences et des techniques est pour une large part celle des ruses permettant d'amener le monde sur cette surface de papier. Alors, oui, l'esprit le domine et le voit. Rien ne peut se cacher, s'obscurcir, se dissimuler.***
>
> ***Bruno Latour, Culture technique, 14, 1985 (cité par Christian Jacob dans L'Empire des cartes, Albin Michel, 1992).***

L'informatique, les outils de la linguistique de corpus ont permis aux linguistes de rester les maîtres de corpus aux dimensions exponentielles. Même chose dans le domaine de la médecine : disséquer une grenouille est une chose (et un corpus), séquencer le génome humain en est une autre. Dans tous ces cas comme dans les **courbes proposées par Google**, le scientifique est parvenu à "ruser" le monde pour user de son corpus.

Et donc ? Nos sociétés de données, nos sociétés d'une **exponentielle et inconcevable immensité de données**, nourrissent en permanence des monstres calculatoires et industriels (voir les textes d'Hervé Le Crosnier sur le sujet, **là** ou **là**) qui, dans certains domaines, sont en passe d'être les seuls capables de circonscrire des corpus qui relèvent, pourtant, du bien commun. Aujourd'hui déjà la génomique, demain peut-être la linguistique, après demain qui sait, les traits culturels ? **Culturonomics.** Le génome de la culture.

S'il est vrai, **comme le remarque Jean Véronis dans son billet** que "*la biologie et le traitement des langues partagent beaucoup de choses du côté des algorithmes et des mathématiques*", je pense que le choix terminologique de Google dépasse, de loin, la seule interdisciplinarité ; Culturonomics : dans l'histoire de Google comme dans ses liens **les plus intimes**, la culture et le génome sont les deux brins d'un même ADN fondateur.

Moralité. Celui qui peut **dire que la vie l'emporte sur la mort** ne doit jamais se retrouver en situation d'être le seul à pouvoir le dire. Ou à prétendre le contraire. Ou à ne pas le dire. Il est de notre responsabilité collective d'y veiller. Habemus corpus. Ceci est notre corp(u)s.

>> Article initialement publié sur **Affordance**

Retrouvez notre dossier :

Petite histoire de la géologie en quelques mots

La politique, le sexe et Dieu dans Google Books

>> Illustrations FlickR CC : **Calamity Meg**, **J.Salmoral**

>> Illustration de Une FlickR CC : **stefernie**

1 ping

Les tweets qui mentionnent Culturonomics: juste une question de corpus? » Article » OWNI, Digital Journalism -- Topsy.com le 11 janvier 2011 - 13:03

[...] Ce billet était mentionné sur Twitter par Christine Genin, Mr mécénat et des autres. Mr mécénat a dit: Numérisation des livres, faut-il un habeas corpus du numérique ? http://bit.ly/fiUa9J #owni [...]