Comment constituer un corpus de documents écrits au sujet d'une controverse sociale ?

Il faut se limiter à des textes, indépendamment et vérifier évidemment leur source, c'est-à-dire que seuls les documents font foi. Alors de nos jours, chercher des documents, ça veut dire chercher des documents sur le Web en général. Alors mon point de vue, là, c'est de rappeler la nécessité d'un point de vue, c'est-à-dire que pour recueillir des documents, il s'agit pas d'accumuler des gigas, il s'agit de bien savoir ce qu'on veut de façon à avoir les bons textes.

Je donne un exemple, si vous entrez dans un moteur de recherche, vous voulez arrêter de fumer, vous entrez "arrêter de fumer", qu'est-ce qui va vous sortir, il va vous sortir des liens commerciaux de l'industrie pharmaceutique qui vont vous proposer des substituts nicotiniques, des pages d'entreprises qui vont vous proposer des séminaires expérientiels, je cite, des sites alternatifs enfin des gourous divers, et cetera et des sites d'hygiène. Si vous entrez comme requête "tabagisme" à ce moment-là vous avez des réponses complètement différentes. Et ses réponses complètement différentes sont beaucoup plus sérieuses pour arrêter de fumer parce qu'elles émanent du Comité national contre le tabagisme et finalement il y a très peu de liens commerciaux là dessus, ce ne sont pas du tout les mêmes liens qui sont présentés.

Un autre exemple, si vous voulez faire un recueil de sites racistes, vous tapez "Bougnoules", vous n'avez que des sites antiracistes parce que le mot bougnoule n'est employé que par les antiracistes tout simplement parce que ça correspond à l'idée que les antiracistes se font des racistes, mais les racistes ont depuis longtemps changé de langage, Bougnoule c'était employé au temps de la Guerre d'Algérie. Donc les choses ne sont pas si simples, ce que je veux dire là, c'est qu'il faut bien préparer ce qu'on veut et être capable de changer ses modes d'interrogation de façon à avoir, disons, des textes pertinents par rapport à la requête. Parce que si vous ne les avez pas, vos résultats seront si on peut dire pollués par les mauvais textes ou par des textes non pertinents.

Alors une fois que vous avez recueilli des textes, il faut savoir que le sens de votre texte va dépendre du contraste avec d'autres textes. C'est un point de vue général en sémantique. Moi je suis sémanticien, la sémantique, c'est l'étude du sens, un mot, ça se définit par opposition à un autre, assis par rapport à debout, noir par rapport à blanc, etc, etc. Et ça veut dire que le sens est fait de différences. Pour savoir ce que dit un texte scientifique, il faut savoir ce que dit un texte antiscientifique, dans le cas d'une polémique, dans le cas d'une controverse, c'est à dire d'une opposition à l'intérieur de la science entre deux hypothèses différentes, c'est un cas différent, je n'en parle pas ici pour le moment.

Par exemple, pour savoir ce que dit un texte, ce qui est caractéristique des textes racistes, il faut les contraster avec les textes antiracistes. Par exemple le mot "homme" est absent des sites racistes. Bon c'est évident quand on le dit, mais encore faut -il le prouver. Le mot "Hitler" est partout présent dans les sites antiracistes, il est presque absent dans les sites racistes. Donc rien que le contraste des corpus vous permet de trouver des caractères auxquels vous n'auriez pas pensé. Alors ça, c'est un point de vue de méthode majeure parce que si vous allez sur le Web, vous prenez des gigas en fonction d'un mot clé que vous aurez choisi peut être intuitivement, vous risquez de ne pas faire émerger ceux qui vous intéressent.

UVED
Université Virtuelle Environnement & Développement Durable

François RASTIER, CNRS

Alors la question, une fois que vous avez des textes, c'est de savoir quels sont les critères pour organiser les informations que vous allez extraire. Alors première chose, ces textes appartiennent tous à des discours déterminés, il y a des gens déterminés. Je prends l'exemple d'un texte législatif, un texte législatif sur l'interdiction de fumer, vous allez le retenir si vous travaillez sur l'interdiction, il fait autorité, mais si vous travaillez sur la réception de la loi, il faudra chercher des blogs de gens qui veulent ou ne veulent pas arrêter de fumer et à ce moment-là, ça sera un autre genre qui sera le centre de votre campus. Donc le filtrage du document, il s'agit pas d'accumuler des textes, il s'agit de filtrer des textes, c'est à dire d'éliminer le maximum de textes non pertinents, il ne s'agit pas de chercher des données, il s'agit de pouvoir éliminer des données de façon à avoir le maximum de caractérisation, de contraste entre les textes pertinents.