Quels indices faut-il prendre en compte pour l'étude d'un corpus ?

Pour analyser les textes, maintenant, je ne vais pas vous faire un cours de linguistique. Mais il me semble qu'il faut, en fonction de l'application, en fonction de ce que vous cherchez, il faut pas se fixer sur une seule catégorie, ni sur des mots particuliers, ni même sur une seule catégorie de mots, vous voyez ? Il s'agit pas de chercher seulement les noms propres ou seulement les pronoms. Ce qui organise un texte, c'est justement les corrélations entre ces différentes catégories. Donc ça ne sert à rien de prendre les catégories grammaticales, le verbe, le nom, etc. , et de chercher soit un certain nom, soit disons des variables linguistiques diverses.

Il y a par exemple des variables, comme la longueur moyenne des mots, qui sont très discriminantes. Mais c'est dans aucune grammaire. Je peux vous dire tout de suite si un texte est technique ou pas technique en fonction de la longueur moyenne des mots, vous voyez ? Bon.

Alors là, il y a tout un domaine de recherche que les linguistes eux-mêmes commencent à découvrir. D'autre part, il y a des indices qui sont extrêmement discriminants, et qui ne sont pas dans les grammaires, comme par exemple la typographie, la police de caractères, vous voyez. Les polices des sites racistes sont pas du tout les mêmes que celles des sites antiracistes. Il y a même d'ailleurs des sites un peu extrémistes qui permettent de télécharger obligeamment des polices néogothiques. Et puis bon, mais les codes de couleurs, tout simplement, hein. On reconnaît tout de suite la couleur politique d'un site en fonction de ses couleurs. Enfin bref, rien n'est à négliger si vous voulez faire une analyse différenciée. Et parfois, des indices qu'on pourrait appeler de bas niveau, comme on dit en informatique, sont très discriminants.

J'ai parlé de la typographie et des couleurs, mais il pourrait y avoir par exemple la ponctuation. Un antiraciste n'écrit jamais une phrase en majuscules. Un raciste écrit des phrases en majuscules. Pourquoi ? C'est peut-être l'équivalent du coup de gueule, ou quelque chose comme ça. Mais il faut être attentif à ça, vous voyez. Et c'est pareil pour les textes faussement scientifiques qui en contredisent d'autres, hein. Il y a des indices, par exemple les sites négationnistes, ça existait et ça existe toujours, étaient une caricature des sites scientifiques. C'est à dire, on vous expliquait avec des tableaux de chiffres que c'était impossible de faire tenir tant de personnes dans un dans une chambre à gaz. Et donc, il y a différentes stratégies, vous voyez. Et il faut pas non plus en rester à l'idée que l'apparence du texte crée sa scientificité. Vous avez des textes qui sont complètement antiscientifiques et qui imitent à merveille un texte scientifique. D'ailleurs, les entreprises notamment de l'industrie chimique ne se prive pas d'en fabriquer, de façon à obscurcir la réalité quand, la vérité scientifique, quand il y a un problème pour elles de rentabilité. Alors, donc la principale recherche, c'est la recherche de thèmes. Mais un thème, c'est pas nécessairement un mot, hein. Il faut dépasser cette question du mot-clé. Souvent, c'est des associations de mots, c'est lié à des ponctuations particulières, à des positions particulières dans le texte, etc. Et de tout cela, on peut tenir compte, en fonction notamment des genres selon qu'il s'agit d'un document institutionnel, d'un article de presse, d'une information commerciale, de pages personnelles, etc. Et donc, chaque genre à son répertoire de thèmes.

Si vous utilisez des logiciels pour étudier vos textes, il faut bien vous dire que les résultats de vos logiciels ne sont que des... Que ce ne sont que des sorties, et non pas des résultats. C'est-à-dire qu'elles ont à être interprétées. C'est pas parce que vous aurez un chiffrage quelconque, a fortiori des chiffrages de fréquences hein, les grosses fréquences ne veulent rien dire à personne. Parce qu'elles sont partout les mêmes, dans tous les textes. C'est pas parce que vous aurez un chiffrage de fréquences que ça aura la moindre pertinence. Donc il faut rester dans les fréquences moyennes, voire dans les fréquences basses. Et moi, je m'intéresse beaucoup aux fréquences 1 et aux fréquences 0, c'est à dire les mots absents. Parce que derrière ça, je vais y venir tout à l'heure, il y a une notion du langage, vous voyez ? Le langage sert tout aussi bien à désigner qu'à cacher, hein. Il y a un spécialiste de la politique à Moscou qui disait, à propos d'un président Russe en exercice, il disait "Les mots servent à cacher les phrases". Et bien souvent, on a par exemple dans les textes issus des firmes de l'industrie chimique, à propos de la disparition des abeilles, des mots qui sont ceux des apiculteurs eux-mêmes, mais elles sont réutilisées et restructurées de façon tout à fait différente. Donc on al l'air d'abonder dans votre sens, et on dit tout à fait le contraire. Je donnerai un exemple tout à l'heure. Bon

L'idée donc, c'est qu'il faut pas se fixer sur des indices simples, se limiter à des mots qu'on va appeler des indices. Il faut chercher notamment les mots absents. Par exemple, le dernier rapport de l'ANSES sur la disparition des populations d'abeilles ne comporte pas, sur 30 pages, le mot "Pesticide". Le mot n'est pas prononcé. Donc il y a un substitut qui va être environnemental. Mais on ne sait pas dans l'environnement, ce qui est important dans cette affaire.

Et pour les firmes qui produisent des cigarettes, c'est la même chose, vous voyez. On parle de facteurs environnementaux. Et l'environnement, bah ça peut être les fumées diverses, mais pas nécessairement le tabac. Donc il faut être sensible à ce qui n'est pas dit. Alors que dans un texte scientifique, on essaye de tout expliciter, vous voyez ? C'est une stratégie différente.