Sociologie et Lexicométrie
Approches lexicométriques comparées en sociologie

François Leimdorfer
Laboratoire Printemps
CNRS/Université de Versailles/Saint-Quentin-en-Yvelines
francois.leimdorfer@uvsq.fr

I. Le recours à l'analyse informatisée de textes est de plus en plus fréquent en sociologie. On peut même dire que c'est une des voies d'entrée dans la problématique du langage pour de nombreux sociologues (en plus de l'analyse des catégorisations). D'autant que le matériau de recherche en sociologie est très souvent, sinon majoritairement, formé de langage : entretiens, questionnaires à questions ouvertes, textes, documents, échanges langagiers divers, et que les corpus réunis sont de plus en plus abondants. (l'importance des corpus est souvent cité comme raison à l'utilisation de logiciels). De plus, il est fréquent que les sociologues (des « sociologues pressés ») attendent des instruments d'analyse des résultats directement exploitables, par exemple sous forme de catégories de « contenu ». (Or selon moi, les logiciels lexicométriques sont plus des instruments de recherche que des instruments de résultats.)

Cependant, il est fréquent en sociologie, hélas, que l'on considère que le langage est « transparent », c'est-à-dire que le sens est directement saisissable intuitivement, alors que pour les linguistes, sociolinguistes, pragmaticiens et analystes de discours, celui-ci est problématique, à reconstruire.

Du côté des praticiens de la lexicométrie, il arrive que les pratiques soit virtuoses et très techniques, mais que les interrogations sociologiques, notamment sur le « discours », fassent défaut.

II. Langage et discours en sociologie

Que veut dire considérer le langage et les discours en sociologie et pratiquer la lexicométrie ?

I. La sociologie essaie de décrire les pratiques, les rapports, les mouvements, les activités, les objets, les regroupements, les situations (des et entre humains), etc., soit en tant qu'ils sont récurrents, soit en tant qu'ils ont des effets, dans un espace et une temporalité donnés. Cela renvoie à de la quantité (lorsqu'il y a récurrence), voire institution lorsqu'il y a stabilisation (des rapports, pratiques, situations, etc.), soit à de la qualité (fait unique, événement qui a des effets) et à de la significativité. En matière de discours et de lexicométrie nous sommes bien sûr au cœur d'un rapport entre quantité et représentativité (d'un regroupement, d'une catégorie par exemple) et qualité et significativité (les effets, les places). La lexicométrie est particulièrement adapté à la mise en évidence des récurrences.
II. Dans cette perspective, le langage en général (et le sens) opère un lien entre individus-acteurs ou acteurs collectifs (entre eux) et situations, ici et maintenant, mais aussi (imaginairement) entre passé-présent-futur, entre l'ici et l'ailleurs. Et dans cette perspective ensuite, il faut envisager ce que l'on appelle « discours », c'est-à-dire la-les paroles insérées dans les situations sociales (et non pas le langage-la langue en soi), comme une « activité », c'est-à-dire pas seulement comme un « plan » langagier, ce que l'on a naturellement tendance à faire lorsque le texte est le lieu de l'analyse. Et dans cette perspective enfin, l'étude du discours nous permet de dégager des indices, des indications de rapports sociaux d'une part, de considérer les discours comme construisant ou actualisant des rapports sociaux, des places, de manière dynamique d'autre part. Enfin, le discours lui-même étant une activité sociale, on peut rechercher des régularités discursives : des « genres de discours » (au sens de Bakhtine), c'est-à-dire des formes langagières régulières (par exemple les recettes de cuisine), des « registres de discours » réguliers (par exemple le registre de la cuisine). S'agissant des « genres », il faut remarquer qu'ils sont presque toujours issus d'un acte canonique : par exemple un « faire-part », un « débat », une « ordonnance », etc.

S'agissant des « registres » il faut distinguer les registres de discours en situation, dans l'ici et le maintenant, et les registres socialement plus ou moins stabilisés (comme par exemple les grands registres disciplinaires, le droit, la médecine, l'économie, le religieux, etc.). Registres et genres sont articulés, dans la mesure où, au sein d'un registre, certains genres sont privilégiés. Michel Foucault avait également évoqué les « formations discursives », espaces de déploiement socio-historiques de discours (par exemple l'espace des discours liés à la colonisation).

III. L'approche sociologique des discours – en sus d'une analyse des situations d'interlocution elles-mêmes - peut se déployer selon deux axes méthodologiques : une analyse des énoncés (énonciations, catégorisations, actes, etc.) et des interactions, et une analyse statistique du vocabulaire sur des corpus à l'aide de logiciels lexicométriques. Ce dernier cas entraîne, dans le cadre de la sociologie, plusieurs conséquences méthodologiques et théoriques.

a/ Tout d'abord, se centrer sur le texte, et donc sur le langage, gomme la situation d'interlocution elle-même, les relations « vivantes » entre locuteurs, les éléments « extralinguistiques » (mouvements, postures, regards, gestes, etc.).

b/ La mise en « corpus », c'est-à-dire l'accumulation de textes (entretiens, écrits, questionnaires, etc.) à partir de « conditions de production » ou de « réception » des discours jugées analogues (locuteurs, situations, etc.) construit un objet d'étude proprement dit où ces discours sont mis à plat. La constitution d'un corpus relève très souvent d'hypothèses intuitives sociologiques sur un « genre de discours », par exemple une étude sur des dictionnaires, sur des publicités, etc. Ces discours sont bien entendu à référer à des situations sociales et des pratiques en réception (consulter un dictionnaire, regarder une publicité) et en production (fabriquer un dictionnaire, une publicité, en vue de...). Ces réserves faites, il est parfaitement légitime de considérer le plan du discours en soi : par exemple les dictionnaires en tant que textes sont un produit social (avec un espace et une temporalité donnés) et fonctionnent de manière « autonome ».

c/ Cela entraîne une mise entre parenthèses des subjectivités des locuteurs et trace ainsi une différence avec une approche psychologique, tout du moins celle qui s'attache aux individualités. En revanche, l'étude du langage (y compris par l'analyse lexicométrique) peut permettre une analyse de la mise en forme langagière de certaines dimensions psychologiques, par exemple celle des émotions (Plantin, 2011).

d/ L'approche sociologique va chercher les régularités du discours, au-delà des postures individuelles, et des écarts à ces régularités. Ces régularités mènent à l'hypothèse que les individus s'investissent dans des « lieux communs » socio-discursifs. Cela n'implique cependant pas, à notre sens, la non-pertinence de la question du « sujet » (*cf. a contrario* les positions d'Althusser ou de Pêcheux : le sujet comme produit de l'idéologie, Leimdorfer, 2011a : 216-217).

e/ La question du locuteur est réintroduite par l'attribution de caractéristiques jugées pertinentes (âge, sexe, études, profession, habitat, etc.). Le locuteur est donc « saisi » à partir de dimensions sociologiques. Les énoncés des locuteurs de même caractéristique (de même sexe par exemple) sont agrégés dans une même classe. Ajoutons que de nombreux discours sont sans locuteur identifiable (lois, dictons, formules, légendes, etc.).

f/ Il faut cependant rappeler une différence essentielle entre « locuteur » et « énonciateur ». Le locuteur est une personne « réelle » pouvant être caractérisée par des catégories sociologiques (ou autres), l' « énonciateur » est le sujet (certes relié à une personne réelle) qui énonce le discours à partir d'une place. Ainsi un écrivain-homme-contemporain d'âge moyen peut parfaitement écrire une narration supposée énoncée par une femme du siècle dernier, jeune ou âgée. De même énoncer une vérité universelle (« l'eau bout à cent degrés ») est indépendant des caractéristiques du locuteur : seule compte la place occupée d'énonciation de discours à prétention universelle. La différence locuteur/ énonciateur à des conséquences importantes dans l'utilisation croisée de *Lexico* et d'*Alceste*.

1. **L'analyse statistique du vocabulaire d'un corpus**, à l'aide de l'informatique, constitue une bonne approche des régularités d'un corpus, surtout si celui-ci est volumineux. Deux logiciels sont ici privilégiés : *Lexico* (A. Salem, 1988, 1994) et *Alceste* (M. Reinert, 1986, 1993). Ces deux logiciels sont construits sur des bases mathématiques analogues (Benzécri) et ont le grand intérêt, de notre point de vue, de ne pas faire de pré-catégorisations thématiques ou syntaxiques (ils ne travaillent que par opérations mathématiques sur le texte), contrairement à « l'analyse de contenu » classique (thématique par exemple) ou certains autres logiciels textuels. Ce dernier type d'analyse avait été critiqué naguère par Michel Pêcheux : les catégories thématiques d' analyse, quel que soit le talent de l'analyste, risquent de reprendre des catégories sociales préexistantes dans la société, et donc de refléter des éléments idéologiques, sans faire les écarts d'analyse nécessaires à une approche sociologique (Pêcheux 1969).

   a/ Les deux logiciels font des opérations de base communes (et maintenant classiques) : segmentation du texte en « formes », calcul des fréquences et des spécificités, co-occurrences. Ces opérations renvoient plutôt à la linguistique (pour le découpage des formes par exemple les langues dites « analytiques » *versus* «synthétiques », pour les fréquences une distribution en courbe de Zipf, pour les co-occurrences les syntaxes de l'écrit et de l'oral, pour les segments répétés les formules et les locutions figées). Elles ont cependant certaines implications sociologiques, une sorte de *sociologie sous-jacente*.

   S'agissant des *formes*, quelles sont les normes de la transcription du langage oral ? Quelles normes orthographiques et syntaxiques, les normes d'une langue « standard », « scolaire », etc.? Doit-on transcrire « ch'ais pas », « je sais pas », « je ne sais pas » ? Doit-on transcrire les intonations, les accentuations ? Les normes de la langue standard et les choix de transcription vont donc avoir une grande importance dans la constitution de la donnée analysable, ce qui peut être acceptable en sociologie, mais pas en sociolinguistique, en analyse des interactions ou en psychologie.

   S'agissant du calcul des *fréquences* des formes, cela renvoie bien évidemment à une conception quantitative comme constitutive du sens, la répétition (des formes), comme sociologiquement pertinente, par rapport à l'événement, à l'unique ou rare, par exemple des énoncés produits par un acteur institutionnel en situation, avec des effets symboliques et pratiques importants. Il y a là une mise à plat, certes légitime, mais qu'il faut resituer dans le contexte des discours et des situations.

   S'agissant des *co-occurrences* et des *segments répétés*, on peut y voir la constitution sociale de répétition et de circulation de « formules » d'une part (Krieg-Planque 2003, Leimdorfer 2006), de contextes phrastiques où les formes prennent des significations particulières, en dehors des définitions sémantiques préalables possibles, d'autre part. On sait l'importance sociologique des formules, écrites ou orales, pour l'analyse de la circulation discursive.

   b/ *Lexico* et *Alceste* présentent des différences d'approches intéressantes pour la sociologie.

   Lexico calcule les « spécificités » des formes simples ou composées, en fonction des caractéristiques attribuées aux unités de discours (locuteur individuel ou institutionnel, type de texte, années, lieux, etc.). Il s'agit de déterminer si des formes se répartissent de manière non aléatoire en fonction de caractéristiques connues de ces unités de discours (âge, sexe, années, origine institutionnelle, etc.). *Lexico* effectue donc une mise en relation des formes à des variables (par exemple les formes spécifiques des hommes, des femmes). Cette mise en relation est « externe », dans la mesure où elle s'appuie sur une connaissance préalable de caractéristiques jugées pertinentes. Le choix des variables caractéristiques participe bien évidemment d'une conception de ce qui peut être sociologiquement pertinent ou non.

   Mais plus fondamentalement, la question du rapport entre un locuteur, ses caractéristiques et ses énoncés se pose : quelle est la nature du lien entre discours et attributs

sociaux de l'individu ? En analyse de discours, l'énonciateur et le locuteur ne sont pas directement assimilables, le locuteur et son discours ne sont pas réductibles à leurs caractéristiques sociales. Reste que la remarque précédente sur la différence entre locuteur et énonciateur s'applique ici : il s'agit de spécificités attribuées à une origine de locution, à des locuteurs. En aucun cas, et il faut parfois insister auprès des sociologues non familiers de l'analyse langagière et lexicométrique, ce rapport est automatique (ce n'est pas parce que telle forme est spécifique des femmes qu'il n'est pas utilisé par les hommes d'une part, et que d'autre part cela n'autorise en rien à parler d'un langage « féminin »).

Par ailleurs, *Lexico* travaille exclusivement sur les formes telles qu'elles se présentent (non « lemmatisées »), ce qui a une grande importance dans le registre des sciences du langage où il s'agit de rester au plus près des formes attestées, telles qu'elles apparaissent. (D'où également l'importance des normes de la transcription, et donc d'un rapport au « réel » déjà reconstitué). Du point de vue sémantique, cela implique (ce qui est la plupart du temps le cas) que de légères modifications de la forme (singulier-pluriel, conjugaison) modifient la signification. Du point de vue d'une sémantique plus sociologique, on peut considérer que les locuteurs disposent d'un « dictionnaire interne » comme dirait Saussure, et que l'existence d'une norme plus ou moins partagée quant à une signification hors contexte et quelles que soient les variations de la forme existe (voir plus loin la lemmatisation).

c/ *Alceste* divise le corpus en segments de longueur déterminée. Les énoncés ne comprenant aucun ou peu de vocables communs sont regroupés et opposés en classes différentes. Le résultat de cette opération met en revanche en évidence des classes qui comportent des termes communs qui leur sont statistiquement spécifiques. Ces classes sont opposées et hiérarchisées (classification descendante hiérarchique). Le logiciel effectue ainsi une partition « interne » du corpus en fonction d'une similitude de vocabulaire des segments. Cette partition nous approche de la conception des « registres » énoncée plus haut : registres internes au corpus dont on pourra faire l'hypothèse (à vérifier) qu'ils ont une extension sociale plus large que celle du corpus précis analysé. De même les termes spécifiques peuvent être une approche d'un genre associé à ce registre.

Ajoutons la question de la « lemmatisation », c'est-à-dire la réduction d'un vocable à une forme canonique qui gomme la conjugaison, le pluriel, le féminin, que le programme standard d'*Alceste* effectue, afin d'aboutir une concentration de vocables. La lemmatisation oriente l'analyse plus vers la signification des « mots pleins » (un renvoi à un objet ou acte réel ou imaginaire du monde) et moins vers celle des formes. C'est une conception de la sémantique différente de celle de *Lexico* (sans être exclusive, en sociologie, l'une de l'autre).

Les caractéristiques connues des énoncés (locuteur, âge, sexe, entretien, etc.) sont attribués après coup comme spécifiques de ces classes. Ici, et c'est important sociologiquement, le classement se fait d'abord sur l'énonciation, puis vers la locution, respectant ainsi la différence locuteur/énonciateur. De ce point de vue, *Alceste*, en constituant d'abord des ensembles de segments de phrases (les classes), puis en les rapportant après coup aux caractéristiques sociologiques, permet de relativiser le rapport entre locuteur et discours.

d/ Les résultats obtenus par les deux logiciels, sans être radicalement différents, montrent deux points de vue sur le corpus. *Lexico* analyse les spécificités du corpus en fonction de variables externes, *Alceste* en fonction d'une similitude de vocabulaire interne (voir exemples plus loin).

e/ Les logiciels ne prennent pas en compte les relations syntaxiques entre énoncés, les interactions (question-réponse par exemple, Achard, 1991, Leimdorfer, 2011b), les relations à la situation réelle d'interlocution, même si on peut récupérer les relations syntaxiques par les tableaux de co-occurrences et certaines marques d'interlocution. En revanche, les calculs se faisant à partir du corpus, ce dernier est la *totalité de référence* statistique et protège d'une conception *a priori* des usages de vocables « dans la société » (d'où l'importance de la constitution du corpus). La possibilité que les logiciels procurent

d'une analyse des formes dans leur environnement phrastique (analyse des co-occurrences) protège également d'une conception *a priori* des significations hors contexte. Le corpus *est* le contexte de référence.

2. **Les résultats** de :
   a/ *Lexico* permettent de mettre en relation des formes (des mots) et des locuteurs (ou des entretiens, des périodes, etc.) et leurs caractéristiques. Il convient cependant de rester prudent : le logiciel n'effectuant pas de lemmatisation, l'interprétation est souvent délicate : pour quelle raison telle forme est-elle spécifique au pluriel et non au singulier ? S'agit-il d'une différence d'emploi et de signification, comme celle qu'on peut supposer par exemple entre « le peuple » et « les peuples » ? Un aller-retour entre résultats et corpus est toujours nécessaire, le logiciel étant un instrument heuristique précieux. Il convient également d'être prudent avec les spécificités mesurées. Ainsi dans notre analyse des réponses aux questions ouvertes de médecins hospitaliers sur le goût et la pénibilité au travail (Estryn, Leimdorfer, Picot, 2010), les femmes médecins manifestent spécifiquement un penchant pour les relations humaines (« contacts », « relations ») ; il n'en reste pas moins que si ces termes ne sont pas spécifiques des réponses masculines, cet aspect de leur travail est (quantitativement) important pour les hommes médecins. Il faut donc toujours se garder d'une attribution mécanique d'énoncés à des types de locuteurs.

   b/ *Alceste* (« algorithme des lexèmes co-occurrents ») permet de constituer des classes d'énoncés et de formes. Ces classes montrent les apparentements, dans le corpus, entre mots (substantifs, adjectifs, verbes) à partir desquels on peut reconstituer des « mondes lexicaux » (Reinert, 1993), des espaces sémantiques. La division entre classes et leurs liens dessinent une répartition des discours par rapport à un objet donné (par exemple les réponses à une question ouverte). L'analyse de ces classes permet d'esquisser des « espaces de points de vue discursifs » sur un objet (Leimdorfer, 2009). Par exemple, s'agissant des médecins hospitaliers et de leurs sentiments de pénibilité au travail, une des classes porte sur la charge de travail (astreintes, gardes) et le rapport au temps libre (familial, vacances, repos, week-end). Il y a donc un « point de vue » qui différencie le monde professionnel (de l'hôpital) et le monde personnel (du domicile), qui insiste sur la temporalité et les rythmes. Cependant, ces espaces renvoient à un locuteur imaginaire construit qui occuperait ce point de vue à partir d'une place abstraite construite également (à partir des termes spécifiques indiquant : localisation, temporalité, activité, temps, personnes). Il faut ainsi différencier cet espace de points de vue et un espace de locuteurs « réels » avec leurs caractéristiques sociologiques (dans notre dernier exemple, les médecins jeunes, dont de nombreuses femmes, sont spécifiques de la classe). Nous pensons ainsi qu'il est possible, s*ur un objet donné* et avec *Alceste* (et à condition que le corpus soit consistant en qualité et quantité), de dessiner une « sociologie des points de vue ».

3. En guise de **conclusion**, faisons un bref retour sur la question du sujet du discours. Une illustration de la différence entre un locuteur générique et un locuteur réel peut être donnée par l'approche lexicométrique *Alceste,* en mettant en regard un monde lexical porté par ce locuteur générique imaginaire et une liste d'énoncés effectifs. En reprenant notre dernier exemple sur le sentiment de pénibilité au travail des médecins hospitaliers, nous avons une liste de termes parmi les plus spécifiques de la classe et quelques énoncés attestés typiques de cette classe :
   - *monde lexical* calculé par *Alceste* : « garde, heure, journée, jour, nuit, repos, rythme, semaine, sommeil, vie, récupérer, fréquent, physique, travail, astreinte, horaire, stress, weekend »
   - *énoncés effectif*s de la classe (termes spécifiques en italiques) : « les *horaires* décalés, souvent 2 a 3 *gardes* par *semaine*, 24 h d'*affilée*: non physiologique, très *physique* provoque des *troubles* du *sommeil* car *travail* de *nuit* très *irrégulier* mais *fréquent*, dommage que l'on

ne soit pas *travailleur* de *nuit* uniquement par *plage* de 1 *mois*, *horaires* trop lourds »
« le *rythme* des *astreintes* et l'*impossibilité* d'avoir un *repos* de *sécurité* après les *astreintes pénibles* du *weekend*, parfois 4 h de *repos* en 2 *jours* avec *reprise* des activités le lundi. »
« fatigue, privation de *weekends* en raison des *gardes*, les miennes et celles de mon mari, *sacrifice* de la *vie* de *couple*, de famille, *enfants*, des *loisirs* et amis. »

J'aimerais enfin évoquer une question en discussion avec Max Reinert, qui est celle de *l'acte* (de discours) que le logiciel *Alceste* pourrait mettre en évidence. La question de l'acte, des actes, des activités par le discours est une question centrale en sociologie du langage. Mais il est paradoxal de penser qu'un discours mis à plat et traité statistiquement puisse montrer un acte. Peut-être s'agit-il des « traces » d'actes, peut-être s'agit-il d'actes sous-jacents lisibles dans les classes d'*Alceste*, peut-être s'agit-il de la notion de « posture » mise en avant par Pierre Achard (1991), lorsqu'en analysant le rapport question-réponse d'une enquête réalisée auprès des conscrits de retour de la guerre d'Algérie, il remarquait que les réponses à la question de leur « plus mauvais souvenir » se distribuaient en une posture de « témoin », d'« acteur » et de « patient » ?

Références bibliographiques

ACHARD P. 1991. « Une approche discursive des questionnaires : l'exemple d'une enquête pendant la guerre d'Algérie », *Langage et Société,* N° 55, pp. 82-96.

ESTRYN-BEHAR M., LEIMDORFER F., PICOT G., 2010. « Comment des médecins hospitaliers apprécient leurs conditions de travail. Réponses aux questions ouvertes d'une enquête nationale », *Revue Française des Affaires Sociales*, Paris, N°4, p. 27-52.

KRIEG-PLANQUE A., 2003, *Purification ethnique, une formule et son histoire*, CNRS-éditions, Paris, 523 p.

LEBART L., SALEM A. 1988. *Analyse statistique de données textuelles*, Dunod, Paris.

LEBART L., SALEM A. 1994. *Statistique textuelle*, Dunod, Paris.

LEIMDORFER F., SALEM A., 1995. « Usages de la lexicométrie en analyse de discours », *Cahiers des Sciences Humaines de l'Orstom "Hommage à Michel Dieu"*, N° 31-1, pp. 131-143

LEIMDORFER, F. 2006, « "Tu sais, on est en Afrique", essai d'analyse de séquences discursives orales », *Semen* N°21, Presses Universitaires de Franche-Comté, pp. 49-71.

LEIMDORFER, F. 2009. « La contribution de la lexicométrie (Alceste) à une sociologie des points de vue », *Bulletin de Méthodologie Sociologique,* N° 104.

LEIMDORFER F. 2011a. *Les sociologues et le langage*, Éditions de la Maison des Sciences de l'Homme, Paris, 290 p.

LEIMDORFER F. 2011b. « Analyser les interactions dans le discours : Les interactions discursives proches et lointaines et les limites de la situation en discours », *Recherches qualitatives* (revue en ligne), Montréal, 17 p.

PÊCHEUX M., 1969, *Analyse Automatique du Discours*, Dunod, coll. Sc. du comport., Paris, 141 p., pp. 101-102.

PLANTIN C. 2011. *Les bonnes raisons des émotions, principes et méthodes pour l'étude du discours émotionné*, P. Lang, Berne, Suisse.

REINERT M. 1986. « Un logiciel d'analyse lexicale : ALCESTE ». *Les cahiers de l'analyse des données* 4 : 471-484.

REINERT, M. 1993. « Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars ». *Langage et Société* 66 : 5-39.