Une approche pragmatique de l'analyse Alceste

Nikos Kalampalikis, Serge Moscovici

Une approche pragmatique de l'analyse Alceste

Nikos KALAMPALIKIS* et Serge MOSCOVICI**

* Groupe d'Étude des Relations Asymétriques, Institut de psychologie, Université Lumière Lyon 2

** Laboratoire européen de psychologie sociale, Maison des sciences de l'Homme, Paris

Résumé *: L'approche sémantique est très souvent la seule utilisée pour l'interprétation des résultats des analyses automatiques du discours. Dans le cadre de cet article nous nous intéressons à l'un de ces logiciels, Alceste, souvent utilisé par les chercheurs travaillant dans le domaine des représentations sociales. Nous soutenons qu'une approche pragmatique de la communication et du langage est indispensable du point de vue théorique et pratique. À partir d'une illustration empirique, nous soulignons la possibilité de mettre en avant les traces indirectes de la communication dans le vocabulaire, grâce au concept de température informationnelle et, par là même, nous montrons comment obtenir des indices pragmatiques aux côtés d'indices sémantiques obtenus par ce logiciel.*

Mots-clés *: Alceste, communication, pragmatique, analyse du discours, représentations sociales, température informationnelle.*

« Les mots sont les signes des idées ;
traiter de l'ordre des mots est traiter de l'ordre des idées »
Weil (1844, p. 1)

Des logiciels et des usages

Le traitement automatique du discours n'est pas une idée nouvelle, mais sa diffusion parmi les chercheurs est récente. Que cela soit des associations verbales, des interviews, des réponses aux questions ouvertes d'un questionnaire, et ainsi de suite, le recours aux logiciels d'analyse des matériaux lexiques est devenu une pratique courante. Nous nous intéressons ici à l'un d'entre eux, le logiciel Alceste[1] qui rencontre un succès parmi les psychologues sociaux, en France et à l'étranger, et les chercheurs qui s'intéressent à la théorie des représentations sociales *(cf. Lahlou, 1998)*. Rappelons, pour commencer, le principe de ce logiciel qui permet, selon son auteur *(cf. Reinert, 1997 ; 1999 ; 2003)*, à travers un ensemble de calculs, de cartographier les principaux lieux communs d'un discours, les mondes lexicaux, qui sont des traces purement sémiotiques inscrites dans la matérialité même du texte. La méthode repose sur un découpage du corpus en fragments de taille relativement analogues, dits « unités de contexte ». Ces mêmes fragments sont ensuite classés statistiquement selon une procédure descendante hiérarchique. À partir d'une première distribution des fragments en deux classes les plus différenciées possible - au niveau de leur vocabulaire spécifique - la plus grande d'entre elles se voit re-distribuée à nouveau jusqu'à ce qu'elle se divise en deux. Cette double opération de distribution/classification continue jusqu'à obtenir un nombre stable de classes. L'objectif de cette classification descendante hiérarchique étant la répartition des énoncés en classes marquées par le contraste de leur vocabulaire *(Kalampalikis, 2003)*.

Si cette description est valable, et ne dépend pas entièrement, comme il se pourrait aussi bien, de notre attitude vis-à-vis des logiciels et de certains jugements sur l'emploi des statistiques, alors l'interprétation des résultats ne poserait aucun problème. Or il se pose dans la mesure où on a la certitude de compter et de bien compter, sans être certains d'avoir ce que l'on compte. C'est la question du signe et du sens ou d'une démographie chiffrée qui se présente sous forme de classes lexicales et un réseau de significations à venir. Les chercheurs les utilisent ensemble, comme on le leur conseille. Et il leur reste toutes sortes d'excuses s'ils poursuivent leur but, dont la plus fréquente est la suivante ; l'interprétation des résultats d'Alceste est une interprétation classique d'une analyse de contenu, c'est-à-dire de nature sémantique. À ceci près que la lexicométrie est plus précise et plus rapide. Mais quand nous nous demandons comment cette masse lexique est arrivée d'abord sur une bande magnétique et ensuite dans l'ordinateur nous nous souvenons qu'il y a eu au départ, un dialogue. Prenons l'exemple d'analyse le

Pour toute correspondance relative à cet article, s'adresser à Nikos Kalampalikis, Institut de psychologie, Université Lumière Lyon 2, 5 avenue Pierre Mendès-France, 69676 Bron CEDEX, France ou par courriel <nikos.kalampalikis@univ-lyon2.fr>.

Cette recherche a été financée par le fonds du prix Balzan 2003 attribué à Serge Moscovici.

1. Sigle pour « Analyse des Lexèmes Cooccurrents dans un Ensemble de Segments de Textes ».

plus ordinaire, l'interview individuelle ou collective, qui est loin d'être uniquement une méthode d'extraction du contenu, comme d'extraire le pétrole d'un puits, et pour l'étude de laquelle on pense souvent qu'il suffit de procéder à une sorte de repérage sémantique de plus en plus sophistiqué. Loin de là, l'interview est une forme de communication entre l'interviewer et l'interviewé qui peut être pleine de contrastes et de heurts. Cela peut être en partie le résultat d'attitudes personnelles, mais c'est beaucoup plus le résultat du fait que tous les chercheurs ne voient pas les choses de la même manière de par le rôle que chacun remplit dans la société. La neutralité des interviewers étant un idéal, chaque entretien est processus de communication qui façonne nécessairement ce qui est communiqué. La matière que nous analysons sous forme d'énoncés linguistiques est donc le produit d'une activité intersubjective de communication. Elle est inter-subjective car le langage, en tant qu'activité symbolique de relation à la réalité et de construction de représentations, régule la communication, ajuste la relation des sujets par rapport à eux-mêmes, transcende leur interaction. Sous cet angle, le contenu de cette matière doit être considéré - analysé et interprété par la suite -, avant toute chose, comme extrait d'un événement de communication intersubjective. L'analyse de contenu est restée à l'écart de ces questions, dans les mains des professionnels intéressés par les approches sémantiques. Elle a peut-être payé le prix de cette réclusion en devenant un peu raide. Nous voici donc amenés à nous demander, à propos du logiciel Alceste, si l'approche sémantique est capable des taches que nous lui assignons aujourd'hui. Il se peut qu'une approche pragmatique soit aussi indispensable du point de vue théorique et pratique. Il est possible même que la pragmatique assume mieux dans ce domaine d'analyse de même qu'elle assume déjà quelques uns des rôles autrefois réservés à la sémantique. Et c'est ce que nous voudrions illustrer concrètement dans cet article.

La dimension pragmatique de la communication

Moscovici *(1994)* a amplement souligné l'importance capitale de l'étude de la dimension pragmatique de la communication pour le champ des représentations sociales. Cependant, force est de constater que la place de la pragmatique continue à rester marginale dans ce domaine d'études. Pourtant, cet auteur avait dressé de manière claire et systématique l'inventaire dans lequel un tel intérêt pourrait s'inscrire. Il avait notamment insisté sur une caractéristique essentielle des représentations sociales, au niveau de leur partage social et de leur contenu sémantique, qui malheureusement continue à être prise *for granded* dans nombre d'études dans ce domaine. Il s'agit du fait que les représentations ne sont que *partiellement* distribuées et non pas *complètement* partagées et qu'elles ne sont que *partiellement* liées au contenu de leur expression sémantique.

Autrement dit, la communication d'un message ne coïncide pas exactement – ni uniquement – avec son expression linguistique. Les représentations ont la qualité d'être des présuppositions dont l'explicitation semble souvent superflue, inutile, une sorte de pléonasme, dans nos conversations quotidiennes. Préserver le consensus, même le faux consensus, semble être une stratégie de communication qui garantit la paix dans nos relations sociales. Mettre en cause ce consensus « mou », faire la lumière de nos divergences sur les non-dits de ces présuppositions c'est arrêter par là même leur mouvement « flottant », selon l'expression weberienne, c'est-à-dire les limites de leur partage. Cette hypothèse théorique permet entre outre de rompre avec une généralisation qui tend à devenir un axiome dans la psychologie des représentations sociales, à savoir que « tout est représentation » ou encore, que « la représentation n'est que du discours », ou même, que toute communication implique *sine qua non* une ou des représentations. Justement, tout n'est pas représentation, car tout n'a pas besoin de l'être. La frontière du contexte d'une représentation – et *a fortiori* de son étude – ses limites, ne sont autres que la pertinence sociale de son activation, l'empreinte de sa communication, d'échange, d'évolution et de transformation de ses significations.

Dans le cadre de cet article, nous voulons insister sur la possibilité de mettre en avant les traces indirectes de la communication dans le vocabulaire et par là même montrer comment obtenir des indices pragmatiques aux côtés des indices sémantiques obtenus par le logiciel Alceste. Mais pour ce faire, il nous faut d'abord aborder la nature statistique de la production du lexique en général et sa mesure grâce, en particulier, au concept de *température informationnelle*. Cela va sans dire, mais peut-être mieux en le disant, notre approche dans cet article est exclusivement pragmatique. Et partant, nous n'attacherons pas un intérêt particulier au contenu sémantique de notre corpus.

Une loi étonnante

Quand on considère même de la façon la plus rigoureuse la science moderne, il est difficile de mettre en doute que la découverte d'un effet inattendu, ainsi les rayons x en médecine, le groupe polarisation en psychologie sociale ou la découverte d'une régularité empirique, ainsi la loi de Zipf *(1949, 1974)*, marquent d'une certaine manière un progrès par rapport aux observations existantes. Et ce d'autant plus qu'ils ont un étrange air de simplicité. Il est évident que ce genre d'effets ou de régularités retiennent l'attention surtout à un moment où d'autres nouveautés se produisent, comme ce fut le cas au moment où la cybernétique a fasciné toute une génération par le concept de *feedback* et la théorie mathématique de l'information *(cf. Shannon, 1948)*. Chacun le sait, mais il est important de rappeler que du point de vue de cette théorie, le couple information et redondance est assemblé en un axiome : *plus d'information il y a dans un message, moins de répétitions il y a ; moins de répétitions il y a dans un message, plus d'information il y a.* Considérez une séquence binaire de nombres qui est aléatoire du point de vue statistique. Dans une séquence de ce genre on cherche un pattern, mais on n'en trouve aucun. On aurait pensé autrefois qu'il s'agit d'un cas de désordre complet ou d'absence de forme. Mais ce n'est pas le cas : la théorie de l'information considère que c'est justement une séquence de ce genre qui contient la plus grande quantité d'information.

Un premier diagnostic de cette relation étroite entre communication et langage a été réalisé dès 1932 par George Kingsley Zipf à partir d'une observation empirique relative à la dynamique des mots. Par l'étude des manifestations objectives du flux du discours, Zipf a démontré certaines régularités statistiques du langage. Disons que sous sa forme la plus simple, la « loi de Zipf » a établi que le produit « rang x fréquence » de chaque élément d'un texte est à peu près constant. C'est-à-dire que si l'on attribue au mot le plus fréquent d'un texte le rang 1, au mot qui suit après le rang 2 et ainsi de suite, il s'établit entre la fréquence r (ou la fréquence relative p) et le rang r une relation telle qu'en multipliant la fréquence par le rang on obtient une constante. Cette loi, donne à voir une relation mathématique stable d'ajustement des mots dans un langage donné testé empiriquement à travers l'étude de divers matériaux, comme des œuvres littéraires (par exemple l'Ulysse de Joyce, l'Iliade d'Homère, des pièces de Plaute), mais également des langues nationales vivantes (par exemple l'allemand, le chinois, l'anglais), ou encore les villes et les mouvements de populations, dans une démarche qui au-delà de son apport statistique sur le vocabulaire a inspiré une nouvelle vision du langage et de la communication humaine.

Ce qui est souvent moins connu dans la formulation de cette loi c'est que Zipf avait supposé qu'elle illustrait un certain équilibre du vocabulaire, donc un ajustement du langage, d'où le nom de *loi harmonique*, caractérisé par deux forces – ou deux économies – opposées, celle de l'unification (réduction de la taille du vocabulaire et des significations en un seul mot) et celle de la diversification (accroissement du vocabulaire par l'attribution d'une seule signification à chaque mot) *(Zipf, 1949)*. Selon ses propres mots : « le degré élevé d'ordre dans la distribution des mots dans le flux du discours indique sans aucun doute une tendance à maintenir un équilibre entre la fréquence, d'une part, et ce que l'on pourrait appeler la variété, d'autre part. (…) La question se pose de la nature de la ou des significations qui mènent automatiquement à cette distribution de fréquence ordonnée. Il est probable que la question ne peut pas être complètement résolue quantitativement, car la ou les significations n'appellent pas la mesure quantitative. Pourtant en isolant les facteurs mesurables, on peut obtenir une connaissance de la nature de la signification et peut-être finalement appréhender un peu de sa nature et de son comportement. » *(Zipf, 1974 ; p. 48)*.

L'hypothèse de la « température informationnelle »

Benoît Mandelbrot, mathématicien français connu pour sa découverte des fractals, a continué le travail de Shannon *(1948)* sur la capacité des canaux de communication et s'est particulièrement intéressé au problème de la distribution des lettres ou symboles linguistiques *(Brillouin, 1956)*. Et c'est lorsqu'il commence à s'intéresser à la distribution des mots du vocabulaire qu'il a été fasciné par la simplicité et la régularité décrites par la loi de Zipf et a essayé de lui donner une formulation mathématique et une explication « physique » *(Mandelbrot, 1954 ; 1957)*. Selon lui, pour mieux expliquer l'effet de cette loi, il vaudrait mieux se baser sur la « variété » dans l'utilisation du vocabulaire, en entendant par là, – comme pour Shannon – le potentiel de surprise, de nouveauté, introduite par chaque nouveau mot. « Notre concept d'explication est voisin de celui de la physique. On dit qu'on a expliqué un fait physique, dont la description est compliquée, ou l'aspect inattendu, lorsqu'on a montré que ce fait est une conséquence, logiquement nécessaire, d'autres faits qui sont plus simples ou plus familiers, mais

Tableau 1 : L'hypothèse de la température informationnelle

	Variété	Richesse	Redondance	
Connaissance	+	+	–	diversité/hétérogénéité
Communication	–	–	+	unité/homogénéité

que l'on ne doit pas considérer comme étant des « causes premières » des faits étudiés. De même, s'il est important de savoir que tout texte « obéit » à certains critères pragmatiques de qualité, *il n'en résulte nullement qu'il ait effectivement été construit pour obéir à ces critères* » *(Mandelbrot, 1954 ; p. 3)*.

Le dénominateur commun de ces trois théories – en apparence étrangères entre elles – c'est précisément leur nature pragmatique. Dans sa théorie, Shannon ne s'occupe pas de la signification à proprement parler, mais des « formes vides ». La loi de Zipf, démontre une propriété insoupçonnée et constante des textes, la relation rang-fréquence, s'appliquant aux « mots-formes », c'est-à-dire, aux différentes formes grammaticales d'une même unité du dictionnaire. La théorie des gaz offre un modèle théorique pertinent qui, de par son histoire et certaines de ces axiomatisations, pourrait, uniquement par analogie, fournir une piste féconde pour l'étude du langage ; selon Mandelbrot, « les lois thermodynamiques macroscopiques effectivement observées pour des grandes masses de gaz sont « les seules auxquelles il faille s'attendre « en partant d'hypothèses assez arbitraires relativement à la forme et au mouvement microscopique des molécules censés former le gaz » *(1957 ; p. 7)*. Autrement dit, mesurer la température du gaz et postuler une théorie expliquant sa thermodynamique revient à élaborer des hypothèses macroscopiques (p.ex. la température, le volume, la pression) qui vont au-delà des simples observations moléculaires (par exemple leur position, leur vitesse, leur composition). Au niveau du langage, Mandelbrot *(1954)* a tenté de proposer des indices de mesure de la « richesse » du vocabulaire d'un texte, sa « température informationnelle ». Pour lui, grande température signifie que « les mots disponibles sont bien employés, les mots rares étant eux-mêmes utilisés avec des fréquences appréciables ». Petite température signifie que « les mots sont mal employés, les mots rares étant extrêmement rares » *(1954 ; p. 24)*. Partant de la notion de « variété-information » chez Shannon (1948), Mandelbrot a évalué la redondance d'un message comme étant un *coût* pour l'encodeur, et une *quantité d'information* pour le décodeur. La température informationnelle, ou le *coût*, pour l'émetteur, décroît quand la redondance des transactions verbales accroît.

Notre hypothèse suivant les principes de la température informationnelle est exposée dans le tableau 1 *(cf. ci-dessus)*.

Afin d'expliquer ce tableau, nous pouvons dire qu'un texte caractérisé par une variété de mots sera plus hétérogène et plus riche au niveau du vocabulaire utilisé ; ce dernier présentera moins de redondances et sera plus apte à la transmission de connaissances. Au contraire, un texte plus homogène, sera moins riche du point de vue de son vocabulaire, présentera plus de redondance, donc moins de variété, et sera moins apte à la transmission de connaissances, quoique plus propice à la transmission et l'utilisation de la communication. Essayons par la suite de tester empiriquement notre hypothèse sur un corpus de discours, une production de communication intersubjective.

Une illustration empirique

Nous avons analysé les données d'une étude portant sur les représentations d'une théorie sociale, le marxisme. Le corpus consiste en entretiens individuels non-directifs en profondeur auprès de cent personnes choisies en fonction de leur appartenance politique et de leur affiliation idéologique[2]. Une analyse quantitative et qualitative de l'ensemble du corpus a été entreprise, afin de dégager les catégories mentales, les systèmes de valeur et les styles argumentaires dont on use pour valider ou invalider la portée théorique et pratique du marxisme. Cette recherche, s'inspirant des travaux initiés par Moscovici *(1961)* sur la psychosociologie de la connaissance, permet plus généralement d'examiner les rapports entre théorie sociale et sens commun, ainsi que l'effet des conjonctures historiques sur

2. Plus précisément, 55 hommes et 45 femmes (âge moyen = 39 ans) ont fait partie de cet échantillon, choisis en fonction de leur appartenance politique ou de leur affiliation idéologique, de leur niveau socio-culturel et de leur localisation (Paris-Province, urbaine-rurale).

Tableau 2 : Les classes lexicales et leurs thématiques

CLASSES	THÉMATIQUE	%
1	Souvenirs de l'initiation au marxisme	16.74 %
2	Vocabulaire ouvrier	17.63 %
3	Le marxisme comme théorie	14.44 %
4	Le marxisme comme courant de pensée	9.48 %
5	Discours autour de Marx et des figures marxistes	8.71 %
6	L'économie marxiste	16.67 %
7	La géographie du communisme	16.33 %

l'acceptation et l'intégration cognitive d'un système de connaissances qui se rapporte au devenir social.

Concernant l'analyse de notre corpus, nous avons opté, dans un premier temps, pour une analyse lexicale informatique, en plusieurs étapes, à l'aide de la méthode Alceste. Dans un premier temps, nous avons effectué une analyse de l'ensemble du corpus[3]. Cette dernière, nous a fournit une structure lexicale stable en six classes issue d'une double classification descendante hiérarchique. Ensuite, pour maîtriser tout biais technique dû à la méthode des entretiens, nous avons voulu connaître l'effet des questions des interviewers sur la structuration du lexique que nous avons obtenu. C'est la raison pour laquelle nous avons procédé à une seconde analyse du corpus des entretiens, mais cette fois-ci, ce dernier consistait uniquement en les réponses des interviewés. Hormis une sous-division supplémentaire lors du second traitement, nous avons constaté que la structure du dendrogramme restait stable. En sachant que le principe de la classification descendante hiérarchique consiste en une répartition progressive des énoncés en classes marquées par le contraste de leur vocabulaire, cette stabilité témoigne d'une répartition quasi identique sur laquelle les questions des interviewers[4] n'ont eu aucun effet significatif. C'est donc sur ce dernier corpus que nous nous sommes concentrés pour la suite de notre analyse. Toute au long de cette dernière, qui est loin d'être définitivement close, nous avons essayé de garder un double regard sur nos données. D'un point de vue global, nous avons tenté de déchiffrer la topographie lexicale de l'ensemble et d'un point de vue local, nous nous sommes attachés à mettre en évidence des propriétés pragmatiques de chaque classe lexicale. Pour mesurer la température informationnelle, différents paramètres ont été calculés :

a. le nombre d'occurrences
b. le nombre de mots différents
c. la type/token ratio[5]
d. les hapax legomena[6]
e. l'indice de Gini[7]
f. les mots ayant le maximum de fréquence
g. les variables externes associées[8]

De la sorte, nous avons considéré les classes lexicales issues de l'analyse Alceste comme des variantes d'un langage qui est celui de notre corpus. Nous nous sommes donc focalisés sur les spécificités locales de chacune d'entre elles eu égard à la thématique de la variante qu'elles représentent. Une thématique qui, précisons-le d'emblée, ne doit pas être assimilée à la sémantique, mais plutôt à une *impression sémantique (Kalampalikis, 2003)* formée aussi bien avant l'analyse (grille d'entretien, lecture des entretiens) qu'après

3. Précisons l'important volume de notre corpus (dû essentiellement à la nature, à la durée, donc à la taille des entretiens) qui compte environ un million d'occurrences (n=925887).

4. D'un point de vue quantitatif, les questions des interviewers occupaient 15,6% du corpus du départ.

5. L'indice type/*token* ratio (TTR) est calculé en mettant en relation les différents mots utilisés (type) et le nombre total des mots émis (*token*). Il s'agit d'une mesure de la diversité lexicale d'inspiration peircienne - introduite par Carroll *(1938)* et améliorée par Johnson *(1944)* – qui a, depuis, fait ses preuves dans de nombreuses recherches quantitatives. Plus la valeur de l'indice TTR est élevé, plus riche et hétérogène est le vocabulaire du texte ; à l'inverse, plus l'indice est bas, et donc proche du zéro, plus la redondance lexicale est importante, donc plus pauvre est le texte du point de vue de la complexité lexicale et de la température informationnelle.

6. Il s'agit des mots utilisés uniquement une seul fois dont l'occurrence est égale à 1.

7. Cet indice de concentration développé par l'économiste Corrado Gini mesure le degré d'inégalité de distributions asymétriques (par exemple les revenus, les richesses etc.) ; pour son usage en statistique lexicale *(cf. Labbé, 1987)*.

8. Comme variables externes, nous avons pris en compte pour l'analyse, le sexe, l'âge, le niveau d'études, l'orientation politique déclarée, la profession et la religion.

(vocabulaire spécifique de chaque classe, χ^2 des énoncés, calcul des segments répétés). Le tableau 2 *(cf. page précédente)* donne à voir de manière synoptique les thématiques par classe ainsi que le poids de chacune d'entre elles sur l'ensemble du corpus.

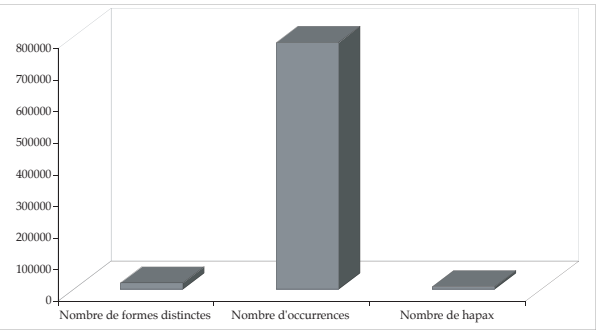
Retenons pour l'instant de ce tableau, que trois des sept classes lexicales contiennent pour l'essentiel la rhétorique marxiste, que cela soit en tant que « théorie » (classe 3), comme « courant de pensée » (classe 4) et comme « système économique » (classe 6). Ceci corrobore également le résultat de la classification descendante hiérarchique, où ces mêmes trois classes sont issues d'une même partition. Ensemble, elles représentent environ 40% du corpus. Les variables externes qui les caractérisent toutes les trois sont le sexe masculin[9], les études supérieures[10] et l'orientation politique commune, proche du parti communiste[11].

Un diagnostic étonnant

De prime à bord, et avant d'entrer dans le détail de l'analyse, nous pouvons remarquer dans le graphique qui suit, un décalage très important et assez étonnant entre le nombre d'occurrences effectif de notre corpus (97%), le nombre de formes distinctes (2%) et le nombre de hapax (1%).

Le graphique 1 traduit d'ores et déjà une donnée de taille caractérisant notre corpus, c'est-à-dire une très importante redondance lexicale. Autrement dit, nos sujets semblent avoir eu recours à beaucoup de mots pour exprimer leurs raisonnements (un peu plus de 800 000 occurrences), sauf qu'il s'avère qu'ils ont souvent, même très souvent, utilisé les mêmes mots afin de l'exprimer. Or, nous savons, au moins, depuis les travaux de Shannon précédemment cités, que, selon les contextes de communication, la redondance peut devenir la face noire de l'information, car plus il y a de la redondance, moins il y a de l'information dans un texte. Inversement, et selon l'hypothèse de la température informationnelle, plus un message est redondant, plus il est propice à l'usage et la transmission de la communication. En sachant que nos « textes » sont issus d'un contexte de communication orale, de langage parlé, qui favorise *a priori* d'autant plus la redondance *(cf. Shannon, 1951 ; Moscovici, 1967)*, ce premier constat demande

Graphique 1 : Aperçu global du lexique du corpus



une certaine relativisation. Néanmoins, étant donné le taux considérable de redondance et la diminution parallèle de la variété et de la richesse du discours, nous pouvons nous demander si l'hypothèse de la température informationnelle *(cf. tableau 1)* ne commence pas d'ores et déjà à prendre corps.

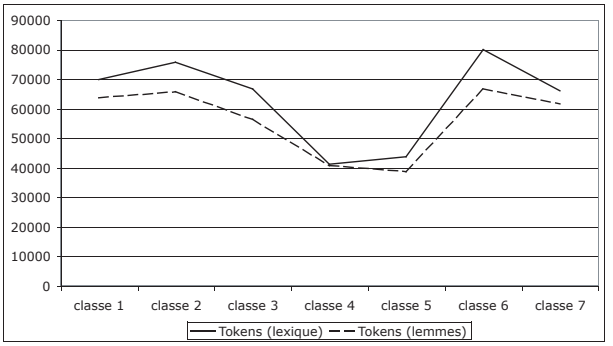
Nous avons donc voulu explorer en détail ce premier aperçu, en calculant l'indice de la type/token ratio (TTR) de l'ensemble. Cette dernière est égale à 0,025[12], résultat qui confirme aussi bien le principe de l'indice évoqué ci-dessus que l'hypothèse de Mandelbrot sur la « basse » température informationnelle. Allons un peu plus loin. Nous sommes visiblement face à un corpus volumineux, mais redondant. Quelle est la distribution de cette redondance globale par classe ? Et à quoi cette dernière correspond eu égard à la thématique de chacune de nos sept classes ? Bref, de quoi parle-t-on quand on répètent souvent les mêmes mots ?

Température lexicale et température lemmatique

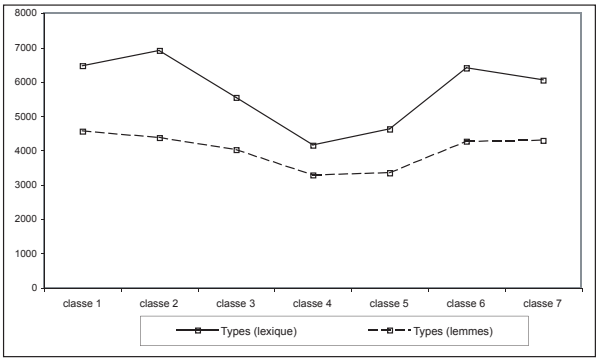
Avant d'y répondre, arrêtons-nous un instant sur un calcul supplémentaire que nous avons introduit. Pour l'économie de l'analyse, Alceste effectue une « lemmatisation », une opération lors de laquelle les différentes formes qu'un mot peut prendre dans un texte (par exemple déclinaisons, singulier-pluriel etc.) sont classées sous une même entrée lexicale, appelée le lemme. Ces lemmes sont regroupés en un dictionnaire de formes réduites à partir duquel nous avons voulu voir comment la redondance se traduisait aussi bien au niveau du lexique (*température*

9. χ^2=18.7 (classe 3), 47.6 (classe 4), 10.4 (classe 6).
10. χ^2=76.9 (classe 3), 36.4 (classe 4), 12.7 (classe 6).
11. χ^2=31.9 (classe 3), 87.4 (classe 4), 9.3 (classe 6).
12. TTR=19686/781214=0,025.

Graphique 2 : Tokens par classe (lexique et lemmes)



Graphique 3 : Types par classe (lexique et lemmes)

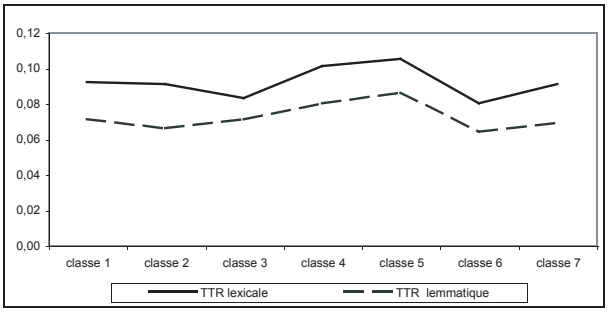


lexicale), qu'à celui des lemmes (*température lemmatique*), cette forme spécifique du lexique, à partir des « tokens » (ou « occurrences »), le nombre total des mots émis *(graphique 2)*, des « types », les différents mots utilisés, *(graphique 3)* et leur ratio *(graphique 4)*.

Le graphique 2 nous donne un aperçu du volume lexical et lemmatique de chacune des classes. On s'aperçoit que la classe 6, qui correspond du point de vue thématique à « l'économie marxiste » est la plus importante vis-à-vis des deux critères de mesure (lexique et lemmes). Autrement dit, lorsque le discours des interviewés se penche sur le thème de l'économie, ils ont tendance à employer un volume conséquent de mots. Ensuite viennent les classes 2 et 1 qui, rappelons le, se réfèrent au vocabulaire ouvrier pour l'une (cl. 2) et au vécu des sujets quant aux souvenirs d'initiation au marxisme pour l'autre (cl. 1)[13]. Plus loin viennent la classe 3 (le marxisme comme théorie) et seulement à la fin, la classe 4 (le marxisme comme courant de pensée).

Là où la tendance s'inverse dans notre corpus, c'est justement au niveau de l'usage d'une variété de mots. Plus précisément, le graphique 3 nous permet de constater qu'au niveau des différents mots utilisés (*types*), et notamment des lemmes, les trois classes lexicales de la rhétorique marxiste (cl. 3, cl. 4, cl. 6) rendent graphiquement la courbe presque droite. Cela signifie que la diversité lexicale est basse et que le lexique varie très peu. Nous sommes peut-être en train d'identifier et de situer une partie de la cause de la redondance générale de notre corpus. Ce qui rend cette observation plus curieuse c'est que la cause se dirige là où on ne l'attendait pas forcément.

Graphique 4 : Type/token ratio (TTR) lexicale et lemmatique par classe



Enfin, le graphique de la ratio lexicale et lemmatique (graphique 4) nous donne un indice supplémentaire qui corrobore notre dernier commentaire. Dans les deux cas de figure, c'est la classe 6 (l'économie marxiste) qui obtient le score le plus bas[14]. Ainsi, malgré le fait qu'elle contient le plus de volume lexical et lemmatique (voir graphique 3), elle se caractérise par une pauvreté lexicale dont la manifestation n'est autre que le haut degré de redondance. À l'inverse, la classe 5 – où on se réfère au personnage de Marx et aux figures marxistes marquantes de l'époque – tout en occupant le moindre poids sur l'ensemble du corpus (8,71% – voir tableau 2), obtient le score le plus haut et s'avère être de loin la classe lexicale la plus variée.

D'autres empreintes lexicales témoignent de nos résultats. Prenons par exemple la fréquence maximum d'une forme lexicale par classe. Une fois de plus, la classe de l'économie marxiste (cl. 6) arrive de loin en tête des six autres (n=3140), suivie de la seconde classe la plus importante de la rhétorique

13. Indice de Gini respectif pour les classes 1-3 : 0,09 ; 0,02 ; 0,01.

14. 0,08 et 0,06 respectivement (indice de Gini : 0,02).

marxiste, la classe 3 (le marxisme comme théorie) (n=2293). En effet, non seulement la classe 6 présente le plus haut degré de redondance de son volume lexical, mais, en plus, elle apparaît avoir le « pic de redondance » le plus important.

En conclusion, et en l'état de actuel de nos analyses, nous sommes face à un corpus lexical extrême, de par son volume, mais aussi de par la distribution de son vocabulaire. Pour l'instant, nous nous sommes cantonnés à prendre la température informationnelle de ce vocabulaire. Notre thermomètre composé de divers indices pragmatiques nous a éclairé aussi bien sur l'état général de la redondance du corpus que sur des zones spécifiques où la redondance bat son plein. Nous sommes donc face à une température informationnelle basse, lexicale et lemmatique, notamment là où on utilise explicitement la rhétorique marxiste, là où le langage de notre corpus devient marxiste. Un langage qui se veut homogène et redondant, peu varié, pauvre du point de vue de sa diversité. Un langage plus propice à la transmission et l'utilisation de la communication qu'à celle de la connaissance.

Discussion

Dans son article sur la dimension pragmatique de la communication Moscovici *(1994)* remarque que *there is something beyond the text* pour insister sur le fait que dans la recherche sur les représentations sociales on doit prendre en considération à la fois la sémantique et l'usage pragmatique du langage. Le concept de température informationnelle nous permet d'étudier ainsi plusieurs aspects : la diversité des vocabulaires lexiques, le degré de prédominance d'un répertoire lexical sur les autres, la différence entre les individus ou les groupes concernant l'usage du corpus lexical disponible. Naturellement, ce n'est pas un hasard si cette relation entre les représentations sociales et la pragmatique est facilement compréhensible. On connaît le rôle que la théorie attribue à la conversation dans la genèse des représentations sociales. Or parmi toutes les modalités d'usage de la parole, la pragmatique concerne « the single most important dynamic context of language use, namely conversation, or face-to-face interactions. The centrality of this functional matrix for language use hardly needs arguing : face-to-face interaction is not only the context for language acquisition, but the only significant kind of language use in many of world's communities, and indeed until relatively recently in all of them »[15] *(Levinson, 1983, p. 44)*. Bien entendu, tout cela doit être nuancé, et il ne faut pas sous-estimer le rôle des modes de relations, des conditions sociales, des intentions pratiques, des possibilités d'expression, dans l'élaboration de représentations *(Culioli, 2002)*. Mais il nous paraît intéressant de souligner, à côté de la conversation, l'intérêt des interactions face-à-face concernant la pragmatique du langage. On sait qu'une des théories les plus significatives dans la communication face-à-face dans les petits groupes a été celle de la pression à l'uniformité *(Festinger, 1950)*, pression visant à réduire l'écart entre le déviant et la majorité. Cette pression à l'uniformité est chose évidente mais les processus de communication qui l'exercent ne l'étaient pas.

Dans une recherche sur les processus de communication et le langage, Moscovici *(1967)* a tenté de montrer quels sont ces processus et découvert que la pression à l'uniformité peut être, dans certaines conditions, une pression à la référence, sur ce dont on parle, sur l'objet de l'interaction, et, dans d'autres conditions, une pression à l'inférence sur les jugements, les opinions des membres d'un groupe. Cette pression agit lorsqu'il est nécessaire de choisir un code commun et de façonner les messages à partir de celui-ci. Elle produit du consensus et se manifeste par la redondance lexicale qu'elle engendre. Redondance qui se traduit par l'usage de répertoires lexicaux communs, des formules admises et socialement actives. Dans ces recherches, le même auteur a montré que la température informationnelle est une bonne mesure du degré de pression à la référence ou à l'inférence, donc de plus ou moins de redondance. On connaît l'importance que cette notion de redondance a acquise dans la théorie de la communication et a été, selon Jakobson « audacieusement redéfinie comme équivalente 'à moins d'entropie relative' ; sous cet aspect elle a fait sa rentrée dans la linguistique ou le concept de redondance embrasse d'une part les moyens pléonastiques en tant qu'ils s'opposent à la concision explicite (la *brevitas* de la rhétorique traditionnelle) et d'autre part ce qui est explicite par opposition à l'ellipse » *(1963, p. 88)*. Il est clair désormais que cette notion pragmatique de température informa-

15. Le contexte dynamique le plus important de l'usage du langage, à savoir la conversation, ou les interactions en face-à-face. La centralité de cette matrice fonctionnelle pour l'usage du langage nécessite à peine d'être discutée : l'interaction en face-à-face est non seulement le contexte pour l'acquisition du langage, mais surtout le seul genre important d'usage du langage dans de nombreuses communautés du monde, et en effet, jusque relativement récemment, dans toute communauté.

tionnelle ne doit pas être négligée non seulement en tant qu'objet d'étude, mais également en tant que méthode, y inclus lorsqu'il s'agit de la famille de logiciels dont Alceste fait partie. Du coup, la question de l'origine de la redondance, pour savoir si elle est due à la recherche d'une représentation commune de l'objet, d'un accord sur ce dont on a l'expression de représentations standardisées ou d'un jugement commun entre ceux qui parlent, devient importante. Si nous rencontrons un tel degré de redondance dans notre corpus, ceci n'est certes pas l'effet du hasard. Si peu de variété combinée à l'importance de son volume illustre peut-être l'effet de la pression à l'inférence. Que cela soit sous forme du lexique ou des lemmes, cette conclusion s'impose particulièrement lorsque les sujets se rapportent au marxisme, comme théorie, comme courant de pensée ou comme système économique. Il n'est pas d'ailleurs exclu que la redondance soit le processus d'une pression à la référence.

On espère que d'autres chercheurs s'intéresseront à ces aspects pragmatiques nouveaux. Cela vient sans doute du fait que, sauf cas extrême, les gens ne parlent pas pour dire quelque chose sans parler pour ne rien dire. Il est rassurant et encourageant de constater que la pragmatique ne concerne pas seulement l'usage des mots mais aussi le plaisir de cet usage. Ou, pour nous exprimer de manière concise, avant de procéder aux comptages des mots, il faut définir le genre de communication dont ils résultent. On est même surpris de voir qu'on ne se donne pas la peine de distinguer les deux modes de communications définissant l'interview, les associations de mots, les dialogues, et ainsi de suite, comme si les mots étaient porteurs de significations intrinsèques. Nous voulons parler de la communication référentielle qui suppose un code commun, des messages verbaux ou verbalisables et un contexte déchiffrable par celui auquel elle est destinée. Ou encore de la communication phatique, qui naît d'un échange abondant d'expressions toutes faites, de représentations ritualisées, voire de dialogues cycliques dont l'unique but est de prolonger la conversation. Elle diminue le coût de chaque transaction informationnelle et elle devient moins propice à la transmission de connaissances signifiantes. Et son répertoire verbal tend sans doute vers l'unification, pour reprendre les termes de Zipf. Mais dès l'instant où un tel degré élevé de redondance permet à la conversation de se poursuivre, au groupe de continuer, une petite cause produit un grand effet. En d'autres mots, dans la compréhension du fait que les exigences sémantiques de la représentation sont indissociables des exigences pragmatiques de la communication, la redondance exprime le collectif, le lien social.

RÉFÉRENCES

– Brillouin L. (1956): *Science and information theory*, New York, Academic Press.

– Carroll J. B. (1938): Diversity of vocabulary and the harmonic series law of word-frequency distribution. *Psychological Record*, Vol. 2, p. 379-386.

– Culioli A. (2002): *Variations sur la linguistique*. Paris, Klincksieck.

– Festinger L. (1950): Informal social communication. *Psychological Review*, Vol. 57, p. 271-282.

– Jakobson R. (1963) : *Essais de linguistique générale*. Paris, Éditions de Minuit.

– Johnson W. (1944): Studies in language behavior. I. A program of research. *Psychological monographs*, Vol. 56, N° 2, p. 1-15.

– Kalampalikis N. (2003): L'apport de la méthode Alceste dans l'étude des représentations sociales. In J.-C. Abric (Dir.), *Méthodes d'étude des représentations sociales*, Paris, Éditions Erès, p. 147-163.

– Labbé D. (1987): Une mesure de la richesse du vocabulaire : l'indice de Gini. *Mots*, 15, p. 171-184.

– Lahlou S. (1998): *Penser, manger*. Paris, Presses Universitaires de France.

– Levinson S. C. (1983): *Pragmatics*. New York, Cambridge University Press.

– Mandelbrot B. (1954): Structure formelle des textes et communication. *Word*, Vol. 10, N° 1, p. 1-27.

– Mandelbrot B. (1957): Linguistique statistique macroscopique. In L. Apostel, B. Mandelbrot et A. Morf (Dirs), *Logique, langage et théorie de l'information*. Paris, Presses Universitaires de France, p. 1-78.

– Moscovici S. (1961): *La psychanalyse, son image et son public*. Paris, Presses Universitaires de France.

– Moscovici S. (1967): Communication processes and the properties of language. In L. Berkowitz (Dir.), *Advances in experimental social psychology.* Vol. 3, New York, Academic Press, p. 225-270.

– Moscovici S. (1994): Social representations and pragmatic communication. *Social Science Information*, Vol. 33, N° 2, p. 163-177.

– Reinert M. (1997): Les « mondes lexicaux » des six numéros de la revue « le surréalisme au service de la révolution ». *Cahiers du centre de recherche sur le surréalisme (Mélusine)*, XVI, p. 270-302.

– Reinert M. (1999): Quelques interrogations à propos de l'« objet » d'une analyse de discours de type statistique et de la réponse «Alceste». *Langage & société*, 90, p. 57-70.

– Reinert M. (2003): Le rôle de la répétition dans la représentation du sens et son approche statistique par la méthode « Alceste ». *Semiotica*, Vol. 147, N°1/4, p. 389-420.

– Shannon C. E. (1948): A Mathematical Theory of Communication. *Bell System Technical Journal*, Vol. 27, p. 379–423, 623–656.

– Shannon C. E. (1951): Prediction and Entropy in Printed English. *Bell System Technical Journal*, Vol. 30, p. 50-64.

– Weil H. (1844): *Question de grammaire générale. De l'ordre des mots dans les langues anciennes comparées aux langues modernes*. Paris, Éditions Crapelet.

– Zipf G. K. (1949/1965): *Human behavior and the principle of least effort. An introduction to human ecology.* New York, Hafner.

– Zipf G. K. (1974): *La psychobiologie du langage*. Paris, Retz.