

Le vocabulaire syndical français à la lumière de l'analyse des données textuelles et de la statistique lexicale.

Mathieu Brugidou
Dominique Labbé

GRETS-EDF, DER, 1 avenue du général de Gaulle 92141 Clamart,
<mathieu.brugidou@edfgdf.fr>

CERAT- IEP (Grenoble II) BP- 48 Grenoble Cedex 9, <cerat@iep.upmf-grenoble.fr>

Abstract :

this article presents a comparative analysis on the leading articles of the French trade-union press in 1996 and 1998. We combine two different methodological approaches: lexical statistics which had been developed by C. Muller and textual data analysis (in particular multivariate methods). In order to highlight common or specific results of these approaches, we analyse union's articles with different softwares (M. Reinert's *Alceste* and D.Labbé's *Lexicométrie*). Some similar results are obtained but each method carries out a lot of specific results. More precisely, multivariate analysis shows the structure of textual data and opens up to several hypotheses. Therefore, lexical analysis offers opportunities of checking these hypotheses and measuring their effects.

Résumé

L'article présente les résultats d'une analyse de discours réalisée sur un corpus d'éditoriaux de la presse syndicale confédérale des trois principales centrales françaises (CGT, CFDT et FO) en 1996 et 1998. Deux approches ont été privilégiées : la statistique lexicale telle qu'elle a été développée par C. Muller et ses disciples et l'analyse des données textuelles. On cherche expérimentalement sur un corpus de textes à dégager les convergences dans les résultats produits et à préciser les spécificités de chaque approche. Ces analyses sont réalisées grâce à différents logiciels (*Alceste* de M. Reinert et *Lexicométrie* de D. Labbé). On observe des convergences réelles entre les deux types de méthodes. L'analyse des données textuelles propose une approche essentiellement exploratoire en mettant en lumière la structure des données. La statistique lexicale permet de s'approfondir et d'enrichir les hypothèses interprétatives issues de la première analyse et de mieux les vérifier empiriquement.

Mots clés : Analyse des données textuelles, *Alceste*, discours syndicaux, lexicométrie, logiciel, statistique lexicale.

L'article présente les résultats d'une analyse de discours réalisée sur un corpus d'éditoriaux de la presse syndicale confédérale des trois principales centrales françaises (CGT, CFDT et FO) en 1996 et 1998.

Deux approches ont été privilégiées : la statistique lexicale telle qu'elle a été développée par C. Muller et ses disciples et l'analyse des données textuelles. La première se propose de construire des modèles et des outils quantitatifs aptes à répondre aux questions posées par la lexicologie et la stylistique traditionnelle. L'analyse des données textuelles applique au langage les outils de la statistique contemporaine et notamment l'analyse multidimensionnelle.

Des logiciels s'inscrivant dans ses deux courants méthodologiques ont été utilisés : les principales analyses des « données textuelles » ont été menées au moyen du logiciel *Alceste*, cependant que l'étude lexicale des textes a été réalisée grâce à la série d'applications logicielles développées par D. Labbé.

Au delà des querelles d'école, on cherche ici expérimentalement à dégager les convergences dans les résultats produits et à préciser les spécificités de chaque approche.

Le champ de l'étude a été défini par quatre facteurs : les choix des organisations syndicales, le genre de discours retenus, les éditoriaux, la sélection d'une série de supports et enfin la période considérée, de 1996 à 1998. Le corpus analysé (118 114 occurrences) comporte plusieurs déséquilibres¹ qui nous rappellent la difficulté à constituer des « corpus représentatifs », c'est-à-dire sans biais.

On observe des convergences réelles entre les deux types d'approches (les principaux clivages sont mis en valeur de manière assez semblable par les deux analyses). C'est pourquoi les deux premières parties de ce texte essaient surtout d'illustrer les résultats spécifiques que permet chacune de ces approches. Enfin, les caractéristiques méthodologiques de chaque approche et les résultats obtenus sur le corpus analysé sont résumés sous forme de tableaux synthétiques.

1. Alceste : la structure des controverses syndicales

L'intérêt de l'approche Alceste (Reinert, 1987) appliquée à l'analyse des discours syndicaux est de proposer une méthodologie qui permette de visualiser et de caractériser la confrontation des discours. En effet, chaque organisation syndicale, à travers les éditoriaux, dessine le contour de la réalité sociale ou du moins cherche à en dresser une représentation. Celle-ci est confrontée à des représentations concurrentes avec qui elle partage des traits communs et dont elle se distingue par ailleurs du fait de l'implantation et de l'idéologie des différentes organisations syndicales. Les discours syndicaux ne s'ignorent jamais totalement. Un « discours » doit toujours être compris comme extrait d'une controverse dont on ne peut que partiellement restituer l'horizon synchronique, - les autres discours syndicaux, patronaux, l'actualité politique, économique etc. - et diachronique. On ne restituera pas ici le détail de l'analyse présentée dans (Brugidou-Labbé, 1999), on précisera seulement, les principales oppositions mises en évidence par la classification et on analysera, à titre d'exemple, une classe d'énoncés. Ces éléments nous permettront de mieux apprécier l'apport de la méthode pour la compréhension des discours syndicaux.

1.1 Deux oppositions

L'analyse statistique permet d'identifier sept sous-ensembles homogènes du point de vue du lexique dans le corpus de textes rassemblés. Ces thèmes sont structurés par deux grandes oppositions visualisées par le graphique factoriel² ci-dessous.

L'axe horizontal du graphique factoriel restitue une première opposition entre des discours centrés sur *la représentation de l'action syndicale* - en distinguant le discours focalisé sur la *mobilisation* et celui sur les *négociations* -, et des discours qui portent sur la représentation des *revendications*, marqués par un important lexique économique.

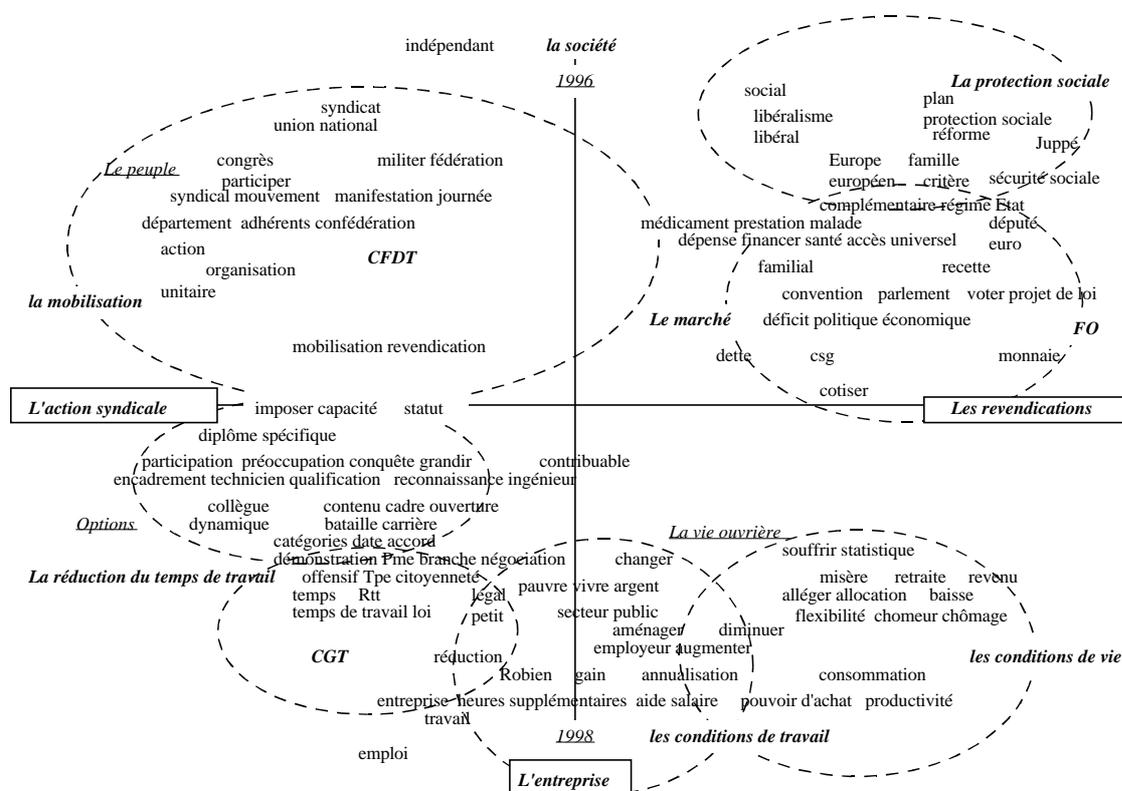
Par ailleurs, on peut opposer deux manières de représenter la réalité sociale : d'une part, on trouve des représentations « concrètes » du monde du travail - dont *l'entreprise* constitue ici le lieu emblématique -, à la fois comme lieu d'action (la négociation) et de revendication (les conditions de travail et les conditions de vie). D'autre part, on recense des représentations plus

¹ Pour une description plus détaillée, voir (Brugidou-Labbé - 1999)

² Le premier axe porte 24% de l'inertie et le second 20%. Les cercles pointillés figurent approximativement les aires recouvertes par les différentes classes. Pour des raisons de lisibilité, on a reporté les formes complètes et non pas les formes réduites à leur racine (stemme) qui sont en fait utilisées par la méthodologie Alceste.

« abstraites », concernant **la société** en général dans ses aspects économiques (le marché), sociaux (la protection sociale) ou politique.

Classification des Unités de Contexte : Analyse factorielle de correspondances



La CGT et la CFDT se situent du côté du pôle de « l'action syndicale », opposés aux éditoriaux de FO qui privilégient nettement la représentation des revendications. On retrouve par ailleurs une distinction nette entre la CGT, qui développe d'avantage le thème des conditions de travail et surtout de conditions de vie, et la CFDT qui privilégie les thèmes du pôle société.

Ces premières remarques doivent être approfondies par une étude systématique des thèmes identifiés. On va s'intéresser plus particulièrement à une des classes d'énoncés qui relève des « discours de l'action ».

1.2 Les discours de l'action : l'exemple de la mobilisation des cadres

L'interprétation des classes identifiées par la méthodologie Alceste s'appuie, sur une caractérisation « externe » de la classe (opposition aux autres classes visualisées par le graphique factoriel) et, sur des critères « internes » (spécificités et énoncés caractéristiques de la classe). Par ailleurs, les variables « illustratives » (mots étoilés ou variables linguistiques : catégorisation des mots-outils etc.) nous permettent d'identifier certaines caractéristiques des textes, notamment leurs auteurs qui sont pour l'essentiel ici issus de la presse CGT. On propose ainsi un « parcours interprétatif » qui s'appuie sur les trois types d'indicateurs mobilisés.

Un premier travail interprétatif consiste à regrouper les spécificités et à les classer en champs lexicaux homogènes.

(/agents/) collègue, encadrement, cadre, ingénieur, technicien, (/identité/) reconnaissance, identité, garantie, catégorie, statut, (/qualification/) qualification, qualifié, diplôme, carrière, (/mobilisation/) unitaire, unité,

mobilisation, convergent, diversité (/participer/) construire, contribuer, participer, débat, exprimer, témoigner (/bataille/) bataille, imposer, offensif, conquête (/fort/) marge, massif, fort, fortement, ampleur (/revendication/) revendication, préoccupation, état d'esprit, mécontentement

Par ailleurs, les énoncés caractéristiques (UCE) permettent de restituer les significations des mots et de tenter de retrouver les logiques discursives sous-jacentes à la classe.

On remarque ainsi qu'il est essentiellement question dans cette classe de la *reconnaissance* du *statut* des cadres et des techniciens : « *Pour les techniciens, nous proposons la conquête de garanties collectives, entre autres l'accès à l'emploi stable la reconnaissance des diplômes, des qualifications, du droit à un déroulement de carrière.* » (CGT -1996)

Ce sont les conséquences de cette *identité* spécifique sur le thème des conditions de travail et de la RTT qui sont analysées : « *Le paradoxe des heures supplémentaires gratuites des cadres et des besoins d'emplois ont motivé particulièrement la participation de nos catégories ingénieurs, de cadres et de techniciens, tout au long du mois de mai.* » (CGT -1998)

Mais plus encore il s'agit de la mobilisation des ingénieurs, cadres et techniciens et de la constitution de ce groupe social à travers la reconnaissance des *collègues*. C'est pourquoi dans cette classe d'énoncés, on note à la fois un suremploi des noms et des mots outils marquant un lien (conjonction « et », adverbe « avec » ou encore possessif « nos »). La mobilisation consiste en effet essentiellement à mettre ensemble et à nommer.

Les textes ici multiplient les références à la *construction* dans l'action du groupe par lui-même. Des expressions comme la *dynamique revendicative*, le *passage du mécontentement à la mobilisation*, voire le *passage à un rapport de force de plus haut niveau* illustrent la volonté de « faire tenir ensemble » des catégories (notamment ingénieurs et techniciens) qui peuvent parfois avoir quelques difficultés à se reconnaître des intérêts communs.

Ces éditoriaux soulignent la difficulté du travail syndical : les *intéressés* doivent être *partie prenante* du collectif en formation sur la base de leur *identité propre*, c'est-à-dire à partir de ce qu'ils se « représentent » de leurs intérêts. Pour cela, il faut que le syndicat n'*assène* pas les objectifs du mouvement social.

Cette dialectique subtile qui fait cohabiter différentes *identités* dans le mouvement social et confond le travail de médiation syndicale et la *construction* du groupe par lui-même n'est possible que par une dénonciation des inégalités économiques et des *attaques patronales*. La prise de conscience du groupe est en effet d'abord la prise de conscience d'un *rapport de force* (cf. *bataille, imposer, offensif, conquête*) dans un premier temps défavorable aux individus qui se trouvent isolés (*mécontentement*) puis favorable dans la *mobilisation* collective (cf. les fortes fréquences dans cette classe des termes comme *massif, fort, fortement, ampleur*).

L'analyse de cette classification permet ainsi de faire apparaître la structure des données lexicales et au delà d'esquisser une typologie des discours syndicaux. L'approche d'Alceste montre clairement les deux grands « objets » du discours syndical : le contenu des revendications et l'action revendicative elle-même³. Elle montre aussi les controverses importantes qui traversent le champ syndical et qui s'expriment par des tentatives de redéfinitions des revendications et par des représentations différentes de l'action syndicale.

³ « Revendication, le mot vient du latin juridique, *rei vindicatio*, réclamation d'une chose. Il désigne à la fois l'acte de demander et le contenu de la demande », (Mouriaux 1983) p. 109.

2. LE DISCOURS SYNDICAL AU CRIBLE DE LA STATISTIQUE LEXICALE

La première étape a consisté à normaliser les graphies puis à lemmatiser les textes (rattachement de chaque forme préalablement normalisée à un lemme et à une catégorie grammaticale). La stabilité de la norme est assurée par le caractère largement automatique des procédures et par l'existence d'un code pour résoudre les difficultés restantes (Labbé 1990a).

Les fichiers lemmatisés permettent une multitude d'analyses statistiques (voir tableau récapitulatif à la fin de cette communication). Nous voudrions ici présenter un aspect intéressant : la comparaison des corpus du point de vue de leur style et de leur vocabulaire.

Puisque la même norme a été suivie dans le dépouillement d'autres corpus, il devient possible de comparer différents vocabulaires.... Certes, il faut comparer ce qui est comparable : un journal avec d'autres journaux et, si possible, la presse syndicale à un corpus de référence représentatif des articles économiques et sociaux, parus à peu près à la même époque, dans l'ensemble de la presse française. Ce corpus de référence est en cours de constitution, en collaboration avec Jean Leselbaum (société Signifier, Paris). Dans la version utilisée cette analyse, le corpus était composé de 436 articles économiques et sociaux, parus entre janvier 1996 et décembre 1998, et tirés au hasard dans le *Monde* (200.000 mots), les *Echos* (200.000 mots) et les hebdomadaires *l'Expansion* et *Capital* (235.000 mots). Soit au total 635 000 mots et 21 550 vocables différents. Naturellement cet ensemble est encore trop petit pour être considéré comme réellement "représentatif". Aussi convient-il de considérer avec prudence le détail des résultats. L'expérience vaut surtout comme esquisse des informations qu'apportera ce type de démarche.

2.1 Un "style syndical" spécifique

Le corpus des éditoriaux de la presse syndicale présente un style bien particulier. En premier lieu, celui-ci se caractérise par une longueur de phrase assez importante et par une complexité relative de celle-ci. L'examen de la densité des catégories grammaticales permet de préciser ces caractéristiques du "style syndical".

Pour les noms propres et les nombres (dates, chiffres), la différence est supérieure à 50%. Or ces deux catégories fournissent une bonne partie de l'ancrage spatial, temporel et social du discours. Leur absence relative dans le discours syndical — surtout dans les textes de la CGT et de FO — signale-t-il une fuite dans les généralités et l'abstraction? Ou bien s'agit-il d'une caractéristique propre au genre "éditorial" ?

L'excédent significatif de verbes (+8,5%) provient pour moitié de l'infinitif (très caractéristique du slogan) mais aussi des verbes fléchis (d'où un excédent important des pronoms personnels et en premier lieu du "nous" et du "on" ainsi que du démonstratif "ce" (dont la fréquence est double de celle attendue). Là encore, on peut se demander s'il s'agit d'une caractéristique du discours syndical ou si cela s'explique par le genre "éditorial".

La complexité de la phrase, une certaine lourdeur, se traduisent notamment par un excédent important de conjonctions (+25%) — de coordination (donc, et, ou ...) comme de subordination — mais aussi par un surplus de pronoms relatifs (que et qui)... Toutefois, la CFDT se distingue par une plus grande sobriété dans l'utilisation des relatives et des conjonctions. Cette organisation semble avoir mieux réussi à se couler dans les canons de la communication contemporaine.

Densité des catégories grammaticales dans le vocabulaire de la presse syndicale comparée au vocabulaire économique et social contemporain

| | Corpus (%) | Syndicats (%) | Syndicats/Corpus |
|------------------------------|------------|---------------|------------------|
| Noms propres | 4,5 | 1,8 | - 59,3 |
| Verbes | 11,5 | 12,4 | + 8,5 |
| Noms communs | 21,7 | 22,5 | + 3,6 |
| Adjectifs | 6,8 | 7,0 | + 2,9 |
| Adjectifs ordinaires | 5,6 | 5,9 | + 6,3 |
| Adjectifs du participe passé | 1,2 | 1,0 | - 13,2 |
| Pronoms | 5,1 | 6,6 | + 30,0 |
| Pronoms personnels | 2,6 | 3,0 | + 15,9 |
| Autres Pronoms | 2,5 | 3,6 | + 44,7 |
| Adverbes | 5,3 | 5,7 | + 7,0 |
| Déterminants | 23,2 | 21,6 | - 6,9 |
| Articles | 14,0 | 15,7 | + 11,9 |
| Nombres | 6,6 | 3,3 | - 50,9 |
| Adjectifs indéfinis | 0,7 | 0,8 | +17,0 |
| Prépositions | 18,4 | 17,9 | - 2,5 |
| Conjonctions | 3,6 | 4,5 | + 24,8 |

Sauf pour les prépositions, les différences de fréquence sont toutes significatives au seuil de 1%.

En revanche, le tableau signale, chez toutes les organisations, la faiblesse relative d'un des traits les plus caractéristiques de cette communication moderne : l'adjectif issu du participe passé (-13%). Ce procédé permet de nominaliser le verbe en effaçant l'auxiliaire : "*reçue* hier par le Premier ministre, la CFDT..." Souvent utilisé pour réaliser des appositions et mettre en valeur le sujet de la phrase, ce procédé donne une expression plus tendue et plus économique qui semble encore assez étrangère au discours syndical.

Au total, ces questions de style ne doivent pas être considérées comme secondaires. On entend souvent dire que les syndicalistes parlent une "langue de bois". Le caractère abstrait et la construction souvent lourde du propos contribuent beaucoup à cette mauvaise réputation... Mais au-delà de ces caractéristiques formelles, la comparaison avec le vocabulaire économique et social contemporain confirme-t-elle ces critiques quand au fond ?

2.2 Un vocabulaire singulier

En premier lieu, le vocabulaire syndical semble peu original : il est, en quasi-totalité, inclus dans le corpus de référence (86,6% des vocables fournissant 98,8% de la surface du texte). Autrement dit, à peine plus de 1% du *texte* syndical est "original". L'essentiel de cette originalité vient du groupe nominal (surtout les noms propres et quelques noms communs ou adjectifs). Ces mots "originaux" sont assez banaux — ils signalent plutôt des "trous" dans le corpus de référence, trous qui devraient en partie se combler avec l'agrandissement du corpus — ou imprévisibles : ainsi, dans un des éditoriaux de FO Hebdo, M. Blondel fait-il une citation latine sur la persévérance "diabolique" d'A. Juppé dans sa volonté de réformer la sécurité sociale, mais l'espérance mathématique du latin dans la vie économique et sociale contemporaine est évidemment infinitésimale !

Le calcul des spécificités (Lafon et Labbé-Labbé, 1997) appliqué aux deux corpus considérés ensemble, met en valeur les singularités du vocabulaire syndical :

— Dans les noms propres, on trouve naturellement les sigles des organisations, les titres des journaux syndicaux ainsi que les organismes sociaux et... la secrétaire générale de la CFDT.

— Les verbes en excès révèlent une des principales caractéristiques du discours syndical : la forte présence de tous les verbes dits "modaux" ou "modalisateurs" ou encore "pseudo-auxiliaires" (Kerbrat-Orecchioni, 1981). En premier lieu, les deux modalités logiques : la possibilité (pouvoir, il est possible de...) et la nécessité (il faut, il est nécessaire de...) et les verbes traduisant les attitudes possibles du sujet envers ce qui est dit : la volonté (vouloir), l'obligation (devoir), la connaissance (savoir). La comparaison avec le corpus de référence montre que ces modalités sont toutes présentes en excédent significatif dans l'ensemble du discours syndical. L'écart le plus fort se rencontre "falloir" et "pouvoir" combinés avec le verbe "être" et les adverbes "ne... pas" dont le suremploi est considérable. Cela donne aux textes syndicaux un caractère extrêmement tendu, voire agressif et moraliste, surtout dans le cas de la CFDT où l'on rencontre le plus fort excédent de "devoir être"...

— Les substantifs suremployés permettent de localiser une autre source de tension dans l'excédent des vocables connotant l'affrontement *action, attaque, brèche, combat, conflit, confrontation, coupe, dégât, dénonciation, destruction, exclusion, fracture, front, gâchis, grève, intervention, lutte, manifestation, mobilisation, offensive, opposition, pression, rapport de force, refus, suppression, sursaut...* mais aussi la tension psychologique : *angoisse, attente, besoin, cri, détresse, espoir, exigence, haine, inquiétude, mécontentement, misère, pauvreté, précarisation, préoccupation, urgence, violence...*

Enfin, ces substantifs spécifiques délimitent le champ d'action des syndicats et soulignent l'ampleur des sujets couverts par les éditoriaux : non seulement *le travail (aménagement, conditions, contrat, durée, temps...)*, les *salaires, pensions, retraites, revenus* mais aussi toute la *protection sociale, l'assurance-maladie, la santé, les allocations, la retraite, l'enseignement*, et même la *politique, le parlement, le gouvernement, les ministres, la loi...*

La préposition "contre" arrive en tête de la liste des mots outils (pronoms, conjonctions, préposition). Sa densité d'utilisation excède des deux tiers celle que laisserait attendre sa fréquence dans le vocabulaire économique et social contemporain.

La liste des spécificités négatives est un peu plus courte car, par construction, l'analyse ne porte que sur les mots les plus usuels. Ceci explique qu'on ne trouve dans les noms propres sous-employés que trois grands pays : *Allemagne, Etats-Unis* et *France* et une région : *l'Asie*. De même apparaissent les prénoms usuels mais non les patronymes (dont aucun n'est assez fréquent dans la base de référence). Aussi courte soit-elle, cette liste confirme ce que nous disions plus haut : le discours syndical est peu ancré dans l'espace. Il ne l'est guère plus dans le temps puisque l'on trouve dans la liste des substantifs significativement sous-employés *an, année, mois* (ainsi que le nom de certains mois)...

La liste des verbes et des substantifs signale le principal manque du discours syndical : le faible intérêt porté aux données de fait. Les syndicalistes utilisent peu *compter, coûter, financer, gérer, investir, payer, vendre...* et les substantifs correspondants : *chiffre, compte, gestion* mais aussi : *acquisition, actif, bénéfice, bourse, capital, crédit, dette... perte... vente*.

Ne comptant pas beaucoup, les éditoriaux utilisent très peu les chiffres ainsi que "mille", "millions", "milliards" et le substantif "franc". Plus fondamentalement, les syndicalistes semblent se désintéresser relativement de *l'entreprise, des sociétés, des prix, de la production, des investissements, des marchés, de la concurrence, des produits, de la rentabilité*, etc.

Au fond, cette confrontation fait ressortir un certain verbalisme, une insistance sur les affrontements et les luttes ; un faible intérêt pour la vie économique et, en premier lieu, pour

la situation des entreprises et des marchés ; et, plus fondamentalement, un discours idéologique peu en rapport avec les dimensions économiques et sociales du temps présent. Naturellement, ces conclusions sont assises sur un corpus encore relativement restreint et dont nous ne pouvons affirmer avec une certitude qu' il est bien représentatif du discours syndical.

3. Deux approches complémentaires

Nous résumons les principales conclusions concernant les outils et les résultats obtenus sur le corpus "discours syndical", dans les tableaux ci-dessous :

Analyse textuelle et Statistique lexicale : caractéristiques de deux approches

| | ANALYSE TEXTUELLE « ALCESTE » | STATISTIQUE LEXICALE « LEXICOMETRIE » |
|-----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Préparation du corpus | Transformation possible des majuscules. | Respect de la graphie, de la ponctuation et de la mise en page (paragraphe) d'origine. |
| Unités | <ul style="list-style-type: none"> - Découpage du texte en Unités de Contexte Elémentaires (choix d'un seuil pour la longueur ou calcul automatique) - Stemmatisation (réduction à la racine) sur la base d'une reconnaissance grammaticale partielle (choix d'un seuil de fréquence ou calcul automatique) - Les mots outils ne sont pas pris en compte dans le plan standard (variables illustratives des classes) <p>Les traitements portent d'une part sur les Unités de contexte et d'autre part sur les unités lexicales stemmatisées. Ils peuvent porter sur les textes (UCI).</p> | <ul style="list-style-type: none"> - Création d' un "fichier-image" comportant, pour chaque mot du texte, sa graphie normalisée, son lemme et sa catégorie grammaticale (reconnaissance automatique et complétée à la main) . <p>Les traitements portent sur l' ensemble des formes normalisées, des lemmes et des catégories grammaticales.</p> |
| Type d'analyses et de traitement | <ul style="list-style-type: none"> - Classification (intersection d'une double classification descendante hiérarchique avec une longueur d'Unité de contexte différente). <p>Une proportion, variable selon la qualité de l'analyse, des UCE est non classée.</p> <p>Les classes sont <i>caractérisées</i> par :</p> <ul style="list-style-type: none"> - des spécificités (qui peuvent être présentées par catégories grammaticales) - des Unités de Contexte caractéristiques - des segments répétés <p>Chaque classe est <i>illustrée</i> par des mots outils et par des catégories grammaticales et sémantiques (noms de lieu, date etc.)</p> <p>Il est possible de réaliser les mêmes traitements (spécificités, UC caractéristiques, segments répétés) par tris croisés sur un mot ou une variable signalétique du texte.</p> | <p>Caractérisation du corpus :</p> <ul style="list-style-type: none"> - distance intertextuelle et classification - non-spécificités du vocabulaire - répartition des vocables - univers lexicaux <p>Caractérisation des parties du corpus</p> <p><i>Stylistique :</i></p> <ul style="list-style-type: none"> - richesse de vocabulaire - propension à diversifier l'expression - spécialisation du vocabulaire - longueur des phrases et structure des phrases - densité des catégories grammaticales <p><i>Thématique :</i></p> <ul style="list-style-type: none"> - segments répétés - spécificités |
| Résultats | Structure des données Caractère herméneutique de l'analyse | Mesure par indices normalisés Caractère cumulatif des résultats |

Résultats obtenus par les deux approches

| ANALYSE TEXTUELLE « ALCESTE » | RESULTATS COMMUNS COMPARAISON | STATISTIQUE LEXICALE « LEXICOMETRIE » |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Visualisation des proximités « locales » entre organisations et mise en lumière du vocabulaire responsable de ces proximités.</p> | <p>Résultat commun : visualisation des proximités entre organisations. Mais absence de mesure des proximités dans la méthodologie Alceste.</p> <p>Pas de caractérisation des corpus entiers dans Alceste.</p> | <p>Proximités entre supports (<i>Vie Ouvrière</i> et <i>Fo Hebdo</i>, <i>Options</i> et <i>Syndicalisme Hebdo</i> pour la RTT, <i>CFDT énergie</i> et <i>Le Peuple</i>).</p> <p>Forte homogénéité du corpus (importance du vocabulaire de l'action etc.)</p> <ul style="list-style-type: none"> - <i>distance intertextuelle</i> - <i>non-spécificité du vocabulaire</i> - <i>répartition des vocables</i> |
| <p>Mise en évidence des thèmes du temps de travail et de l'emploi</p> | <p>Mise en évidence du recouvrement des deux univers mais absence de mesure dans la méthodologie Alceste.</p> <p>Liaison entre les mots mesurée sur la base des UCE dans la méthodologie Alceste.</p> | <p>Prééminence du thème du temps de travail et de l'emploi, fort recouvrement de ces deux univers.</p> <p>Existence d' autres univers lexicaux plus mineurs (mesure des densités et des liens entre leurs constituants).</p> |
| <p>Opposition entre deux types de vocabulaires (voire de discours) : vocabulaire de l'action et celui de la revendication.</p> <p>Opposition entre des discours revendicatifs centrés sur la <i>société</i> et d'autres sur <i>l'entreprise</i>.</p> <p>Champs des controverses entre syndicats :</p> <ul style="list-style-type: none"> - FO et CFDT autour de la protection sociale. <p>Proximité dans le discours de FO entre la protection sociale et le thème du libéralisme.</p> <ul style="list-style-type: none"> - CFDT et CGT autour de la mobilisation et de la négociation (et de la RTT). | <p>Absence d' outils mesurant les styles dans la méthodologie Alceste.</p> <p>Thématiques des différentes organisations syndicales :</p> <ul style="list-style-type: none"> - CFDT (RTT et vocabulaire de l'action et de la vie syndicale). - CGT (vocabulaire de l'action et ses connotations, entreprises publiques et salariés, les conditions de vie, etc.) - FO (sécurité sociale, pouvoir d'achat, dénonciation du libéralisme, conditions de travail...) <p>Absence de module d' analyse factorielle dans les programmes de statistique lexicale. Une vue d' ensemble est possible en transférant les tableaux lexicaux vers un programme d' analyse des données.</p> | <p>Analyse des styles : modèle de la « communication moderne » pour la CFDT, de la « propagande » pour la CGT et de la « rhétorique » pour FO.</p> <p>Thématiques des différentes organisations syndicales et des différents journaux (cf. ci-contre).</p> <p>Existence d' autres thèmes mais dont le réseau lexical est moins dense.</p> |

Conclusions

Au delà d'une réelle convergence des résultats, on voit ainsi se dessiner une relative complémentarité des analyses. L'analyse des données textuelles propose une approche essentiellement exploratoire en mettant en lumière la structure des données. La statistique lexicale nous permet de rendre compte de certains biais dans le corpus, d'approfondir et d'enrichir les hypothèses interprétatives issues de la première analyse et de mieux les vérifier empiriquement.

On notera également que l'approche globale et exploratoire de l'analyse des données textuelles a pour contrepartie le caractère faiblement cumulatif des résultats. Certes, la structure, mise en évidence par les analyses multi-dimensionnelles, peut toujours faire l'objet d'une comparaison qualitative avec des analyses de même type menées sur d'autres corpus mais, à ce jour, la mesure de ces écarts ne peut être réalisée que par les outils de la statistique lexicale.

On voit ainsi de dessiner des protocoles possibles d'études combinant les deux méthodologies : l'approche d'Alceste a des vertus heuristiques importantes, en mettant à jour la structure des données textuelles, elle permet de formuler rapidement des hypothèses interprétatives. Les nombreux outils de la statistique lexicale permettent d'éprouver ces hypothèses et de « solidifier » ces résultats en assurant ainsi leur caractère cumulatif.

Références

- Bergounioux A., Launay M.-F., Mouriaux R., Sueur R., Tournier M. (1982) *La parole syndicale*, PUF
- Bernard-Steinacker C. (1993) « Vingt ans de revendication », Mots, 36
- Blot I. et Le Roux D. (1992) « L' utilisation du logiciel Alceste au département GRETS » EDF, ref. HN-52/92/067
- Brugidou M., Labbé D. (1999), *Le discours syndical, CGT-CFDT-FO en 1996 et 1998*. EDF, ref. HN-52/99/013
- Hetzel A.;-M, Lefevre J.Mouriaux R., Tournier M. (1998) *Le syndicalisme à mots découverts*, dictionnaire de fréquences (1971 - 1990), Syllepse, 352 pages.
- Kerbrat-Orecchioni C. (1981), *L'énonciation de la subjectivité dans le langage*, Paris, A Colin.
- Labbé C., Labbé D. (1994), *Que mesure la spécificité du vocabulaire ?*, Grenoble, CERAT.
- Labbé D., (1990b), *Le vocabulaire de François Mitterrand*, Paris, Presses de la Fondation nationale des sciences politiques.
- Lafon P. (1984), *Dépouillements et statistiques en lexicométrie*, Genève-Paris, Slatkine-Champion.
- Lafon P., Lefevre J., Salem A., Tournier M., (1985), *Le Machinal. Principes d'enregistrement informatique des textes*, Paris, Klincksieck.
- Lalhou S. (1995) « La construction du sens dans l'analyse statistique de données textuelles » EDF, HN-51/95/012.
- Lebart L. Salem A. (1994), *Statistique textuelle*, Paris, Dunod.
- Mouriaux R. (1983) *Les syndicats dans la société Française*, PFNSP.
- Muller C. (1977), *Principes et méthodes de statistique lexicale*, Paris, Larousse.
- Reinert, M. (1987), « Classification descendante hiérarchique et analyse lexicale par contexte : application au corpus des poésies d'Arthur Rimbaud », Bulletin de Méthodologie Sociologique, n°13. 1987.