Le Recours aux corpus discursifs : difficultés et possibilités pratiques

Frédéric Torterat

Département de Langues Étrangères et Communication de l'Université Technique de Construction de Bucarest (UTCB)

UFR des Langues et Cultures Étrangères de l'Université Paris Ouest Nanterre La Défense

Faculté de Langues et Littératures Étrangères de l'Université Pédagogique d'État « Ion Creangă » de Chisinau

Sous la direction de : Mihaela Şt. RĂDULESCU
Bernard DARBORD
Angela SOLCAN

Méthodologie de l'apprentissage de la recherche universitaire

Dans le cadre du projet de coopération scientifique inter-universitaire
Méthodologie de l'apprentissage de la recherche universitaire
L'harmonisation des pratiques académiques
financé par l'Agence universitaire de la Francophonie

EDITURA DIDACTICĂ ŞI PEDAGOGICĂ, R.A.

TROISIÈME PARTIE

L'ÉLABORATION DU TRAVAIL DE FIN D'ÉTUDES

A. ÉTAPES

Chapitre 3

LE RECOURS AUX CORPUS DISCURSIFS : DIFFICULTÉS ET POSSIBILITÉS PRATIQUES

Frédéric TORTERAT[1]

1. Les corpus discursifs : spécificités, approches

Cette étude décrit quelques unes des principales caractéristiques d'un type de corpus parmi les plus sollicités en sciences humaines et sociales, à savoir les *corpus discursifs*. A ce titre, elle résume certaines difficultés méthodologiques qui peuvent se présenter à l'analyste, et les possibilités pratiques propres à ces documents pour le traitement et l'analyse des données. En prenant pour exemple l'approche de deux formes de corpus oraux, à savoir les co-verbalisations de jeunes enfants en contexte pédagogique, et les réponses données par des animateurs de l'environnement lors d'entretiens professionnels, cet article en livre une illustration dans les domaines de la linguistique de l'acquisition d'une part, et d'autre part dans celui de la sociologie des professions.

Rappelons avant tout qu'en sociologie comme en anthropologie culturelle, et bien entendu en linguistique, le recours à des éléments de corpus est d'autant plus répandu qu'il s'avère dans bien des cas incontournable, mais que, dans le cadre de la recherche, les analystes qui font un emploi significatif de ce type de support sont confrontés à des pratiques variées. Celles-ci concernent des questions qui vont de la documentation sollicitée aux commentaires qu'il convient de mettre en œuvre, mais aussi le traitement informatique des données, les variables à prendre en compte, la contextualisation des matériaux, ou encore la mise en commun des approches.

L'une des particularités de la recherche sur corpus discursifs consiste dans le fait qu'elle rassemble opportunément deux dimensions, que l'on peut appeler ontologique et pratique si l'on veut, avec, pour la première, la détermination en particulier des *genres de discours* ou des *formations discursives* qui sont sollicités pour l'analyse. Pour ce qui relève des genres de discours, cela suppose

[1] Maître de conférences en Sciences du langage à l'Université de Nice Sophia Antipolis.

de se positionner sur les contours qu'on entend leur donner. Dans une acception restreinte, on admettra ainsi comme tels, parmi d'autres possibilités, les commentaires journalistiques, les débats parlementaires, les vœux présidentiels ou encore les plaidoyers juridiques. Dans une acception étendue, l'analyse portera par exemple sur les discours urbanistiques (Cf. Bulot et Veschambre, 2000), sociohistoriques, politiques ou philosophiques (Cf. Pêcheux 1990, Branca-Rosoff, 1998).

En pratique, les corpus discursifs sont de plus en plus couramment liés à la collecte de productions orales. Cela explique pourquoi, depuis au moins les années 1990, les corpus oraux ont fait l'objet de plusieurs réflexions critiques, lesquelles rendent compte des multiples questions qui accompagnent le recours à ce type de productions, à savoir que ces questions se révèlent tout aussi plurielles que peuvent être les approches méthodologiques mises en place (Sinclair 2005). Effectivement, outre la construction ou l'extraction du corpus, avec tout ce dont elles s'assortissent en termes de tri (dans les domaines de la collecte, du terrain prospecté, des enregistrements et des éventuelles transcriptions), l'informatisation des éléments pris en compte ne va pas sans poser des difficultés elles aussi spécifiques. Les réponses qu'il convient d'y apporter se révèlent d'autant plus pressantes que si certains logiciels comportent des programmes propres à l'analyse de l'oral (comme Praat, Transcriber et Transana), d'autres s'y prêtent plus indirectement, comme c'est le cas d'Unitex par exemple, et ainsi se pose la question de savoir quels sont les aménagements qui doivent être opérés sur les éléments collectés, avant d'intégrer correctement les éléments de corpus dans le cadre d'un traitement informatique.

Les corpus oraux impliquent en effet des données élémentaires ou « primaires » (les enregistrements) et d'autres, complémentaires, qui renvoient notamment aux transcriptions et aux annotations. Cela étant, leur analyse ne peut passer que par l'écrit, avec l'ensemble des aménagements graphiques des données brutes que cela suppose (Gadet, 2003 ; Dister et Simon, 2008 : 55-59). Outre donc la pratique matérielle des enregistrements, lesquels sont aujourd'hui numérisés pour la plupart, les modifications graphiques doivent présenter un minimum de garanties méthodologiques, sans lesquelles ce ne sont plus des aménagements qui sont effectués, mais une réécriture du document. Dans la mesure, par exemple, où il n'existe pas de correspondances régulières entre les phénomènes prosodiques (comme les contours intonatifs) et la ponctuation graphique (Grobet, 1997 : 89-90 ; Shriberg *et al.* 2000), et qu'il est quelquefois difficile de définir *a priori* quels seront les moments textuels minimaux sur lesquels porteront l'analyse (Avanzi et Horlacher, 2007), les transcriptions de corpus oraux ne sont généralement pas ponctuées en tant que telles, mais comportent plusieurs types d'indications liées à l'oralité, au premier rang

desquelles apparaissent les spécifications des pauses, des répétitions ou des interventions simultanées. Or, si l'écrit se conforme à un certain nombre de prescriptions, comme une disposition textuelle unilinéaire, une faible variation graphique et une ponctuation spécifique, l'oral implique des variations, des disfluences (Blanche-Benveniste *et al.*, 1990), des données para-verbales, une corporéité des productions, voire une part d'inintelligibilité dans certains cas. Par ailleurs, en plus du fait qu'il est confronté, encore aujourd'hui, à certaines faiblesses des corpus précédemment abordés, de leur catalogage et de leur archivage, l'analyste ne peut encore prendre appui, pour encadrer sa propre recherche, que sur des traitements pour le moment assez ponctuels ou rattachés à telle ou telle approche, voire à telle ou telle ressource. Qui plus est, de même que la plupart des corpus à base écrite sont supposés *corrigés*, de même est-il possible qu'un corpus oral soit *redressé*, avec des justifications diverses. Quand par exemple une expression orale n'est pas intelligible ou laisse perplexe (ce qui peut être le cas en linguistique de l'acquisition), les rédacteurs entreprennent soit de reporter une multitranscription (plusieurs possibilités envisagées), soit d'indiquer une incertitude, soit, enfin, de redresser l'élément (ou les éléments) du corpus, autrement dit de modifier ou remplacer la donnée brute pour la rendre accessible.

Les corpus oraux exigent de ce fait une série de marquages en lien avec ce qui fonde leurs spécificités (en marge des éventuels « nettoyages » dont ils sont l'objet), comme ceux qui relèvent de l'intonation, de l'hésitation ou du para-verbal (rires, chuchottements, *etc.*). Au reste, le report de toutes ces indications demeure conditionné par les exigences de ce qu'elles sont appelées à rendre compte. Dans cette vue, si l'absence de ponctuation graphique, mais dans le même temps le marquage des pauses et des disfluences, caractérisent en partie les transcriptions en linguistique de l'acquisition, ceux-ci n'ont pas tout à fait les mêmes significations dans le cas d'entretiens sociologiques, qui peuvent d'ailleurs y contrevenir. On admettra toutefois que, dans tous les cas, ce n'est pas le traitement qui doit conditionner la ressource et l'analyse, mais l'inverse.

Pour illustrer le propos traité, nous présenterons ici deux corpus assez différents : le premier renvoie à des dialogues collectés dans le contexte d'une interaction pédagogique entre une intervenante et des enfants de 30 à 48 mois, lequel support correspond aux besoins de la professionnalisation d'enseignants entrant dans le métier. Dans le deuxième cas, les données du corpus coïncident avec les réponses rassemblées lors d'entretiens semi-directifs d'animateurs de l'environnement, dans le contexte d'une recherche sur la sociologie des professions parrainée par plusieurs organisations, dont l'ADEME[2].

[2] L'ADEME est une agence qui, sur le territoire français, opère dans le domaine de l'environnement et de la maîtrise de l'énergie.

2. Principes méthodologiques, en bref

Les corpus discursifs, parmi d'autres caractéristiques, convoquent le discours comme objet même de l'analyse, et posent de ce fait des questions qui ne s'appliquent pas forcément à tous les corpus. Il appartient en premier temps, à l'analyste, de justifier la typologie sur laquelle il s'appuie, de cerner correctement les « objets discursifs » (Longhi 2007 : 150) dont il se saisit, et notamment quelles informations socio-discursives ou socio-historiques, pour ne citer qu'elles, il prend en compte pour sa recherche.

Dans cet esprit, ce qui vaut dans le domaine des « grands corpus » (qui comptent, pour certains, plusieurs millions de mots), s'applique aussi à ceux qui sont moins denses. Par exemple, il est souvent opportun de prélever dans les données un « moment du corpus », pour reprendre l'expression de Jacques Guilhaumou, à partir d'informations diverses, de manière à s'en saisir à l'occasion d'une analyse linguistique plus aboutie. Le fait d'ailleurs que, depuis ces dernières années, les composantes des corpus soient envisagées comme des ensembles multidimensionnels (Roulet 2002), ou encore des *grandeurs textuelles* (Rastier 2007), plutôt que comme des segments formant des blocs uniformes (Mann et Thompson, 1988 ; Jackiewicz 2002 *inter al.*), est pour le moins révélateur.

Les explications données sur l'approche méthodologique d'un corpus discursif en conditionnent donc significativement la réception. Comme l'explique Pincemin (2007 : 5), « pour que les résultats d'une analyse, d'un traitement sur un corpus prennent valeur, et puissent devenir des éléments de réponse à un questionnement scientifique, il faut une interprétation méthodique des produits du traitement, qui, elle-même, s'appuie nécessairement sur une interprétation / compréhension tant de la procédure d'analyse que de la composition du corpus ». B. Pincemin rappelle de ce fait ce qu'elle estime être les trois principales « composantes » que sont le codage, avec notamment l'emploi d'annotations, le contexte (qui renvoie au matériau linguistique et textuel, mais s'étend aussi dans certains cas à d'autres formes de contextualisation), et l'interprétation, qui s'établit dans bien des cas à partir de « rapprochements et de contrastes » (9)[3].

[3] La possibilité de catégoriser commodément les indications contextuelles en données infra-textuelles (parties ou paragraphes, *etc.*) et supra-textuelles (regroupements textuels, analogies...) est de plus en plus couramment sollicitée.

Les quelques principes méthodologiques que nous pouvons en retenir renvoient tout autant à des questions factuelles et matérielles (terrain ? prise de son ? éventuel support filmique ?, *etc.*), qu'analytiques et opérationnelles. Les questions qu'ils concernent nous paraissent être tout particulièrement les suivantes :

- *quelle représentativité du corpus ?*
- *quelle matérialité ?*
- *quelle productivité du traitement ? quelle éventuelle complémentarité s'ils sont plusieurs ?*
- *quelle démarche analytique envisagée, et pour quels types de conclusions ?*

Ces questions en appellent quelques autres, moins généralistes, mais elles s'assortissent de principes de mise en œuvre qui coïncident avec autant d'éléments de réponse auxquels l'analyste est appelé à donner des contours concrets. La représentativité du corpus renvoie ainsi tout autant au contexte, aux conditions de production ou de collecte, qu'aux formations, aux objets ou aux moments discursifs abordés, de même qu'à la typologie qui s'y rattache. L'approche de ce qui touche à sa matérialité suppose de préciser, outre le nombre de caractères qu'il comporte, si le corpus a fait l'objet d'une construction ou d'une extraction, s'il présente un caractère monologal ou polylogal, délocutif ou interlocutif, ainsi qu'à spécifier l'existence d'une transcription et / ou d'annotations spécifiques, d'aménagements graphiques et / ou de commentaires. Concernant le ou les traitement(s) mis en place, la seule indication des logiciels et des programmes employés ne suffit évidemment pas : les conditions de numérisation des éléments, les variables dégagées, l'interopérabilité que le(s) traitement(s) permet(tent) d'envisager sont des indications souvent indispensables. Quant aux explications fournies sur la démarche analytique, celles-ci font généralement apparaître dans quelle mesure ont été prises en compte la spécificité des données et du (ou des) traitement(s), et si elle reprend suffisamment, par ailleurs, les éléments du corpus à l'appui d'une réflexion critique. Multivariée ou non, multimodale ou unifiée, l'analyse doit dans tous les cas se prêter à une interprétation satisfaisante, quand bien même elle aboutirait à de brèves déductions ou à des conclusions intermédiaires.

L'apport documentaire et les recherches antérieures sur des thématiques similaires ne sont donc pas les seuls éléments qui permettent, à l'analyste, de positionner sa propre démarche. Rappelons néanmoins que l'un des avantages que présente aujourd'hui l'informatisation des données consiste dans leur mutualisation, ce qui implique une accessibilité consentie, quels que soient les supports de diffusion auxquels recourent les équipes de recherche. Cette mutualisation va de la simple mise à disposition d'annuaires à la mise en place

des multipartenariats, en passant par l'organisation de réseaux ou la gestion de plateformes dédiées à tel ou tel type de corpus.

Nous n'insisterons pas sur ces points, déjà traités par de nombreux auteurs. Qu'on nous permette, en revanche, de présenter ici brièvement deux types de formes discursives qui, toutes deux, posent des difficultés en partie analogues, en partie singulières, ainsi que les possibilités de traitement et d'analyse qui leur sont propres.

3. Deux types de formes discursives : Co-verbalisations en apprentissage et entretien professionnel

3.1. Premier exemple : un dialogue pédagogique en linguistique de l'acquisition

Le corpus discursif présenté ci-après a été construit à partir des productions verbales d'enfants de 30 à 48 mois enregistrés dans un cadre polylogal, et en contexte pédagogique. Intitulé *DAM07/Nice*, il correspond à trois instantanés recueillis, pour chacun, auprès de trois groupes d'enfants au cours d'ateliers où il s'agissait de fabriquer sur quelques séances des petites roulottes individuelles avec des emballages fournis par les familles. Faiblement redressés, les éléments de corpus, qui forment un petit ensemble de presque 12000 mots, ont fait l'objet d'une analyse multivariée de manière à dégager les domaines de variabilité interindividuelle parmi les groupes, autrement dit à prendre la mesure de ce qui distingue et ce qui rapproche les enfants, l'objectif étant de prendre en compte leur diversité bien sûr, mais aussi d'organiser la classe aussi opportunément que possible.

Comme nous l'avons suggéré dans les lignes qui précèdent, un corpus de ce type contient des données brutes ou annotées, qui peuvent être décrites par des métadonnées de manière à favoriser leur traitement, avec ce que cela implique en termes de conversion numérique, d'archivage, mais aussi d'interopérabilité. Celles-ci s'accompagnent donc couramment de descriptions et de commentaires qui rendent compte des conditions et du contexte de production et de collecte (ce qui permet notamment de mesurer la représentativité du corpus lui-même). Les éléments dégagés, qui font l'objet soit d'une construction à part entière, soit de l'extraction d'un corpus déjà existant, nécessitent donc d'être documentés (voir Baude, 2006 pour un guide des bonnes conduites). Ce qui accroît par ailleurs la productivité des objets revient à la question de leur

accessibilité : plus ils sont faciles d'accès, plus la démarche mise en œuvre est proprement valorisée (Torterat, 2009).

Dans un cas comme celui-ci, la numérisation des données entraîne une partition (par locuteurs ou par *tours de parole* notamment), avec un balisage informatique qui s'accompagne éventuellement d'un *étiquetage* des éléments. Pour ce qui relève du corpus *DAM07/Nice*, nous avons procédé en deux temps. La première version des données brutes, au format « texte », a d'abord été triée (par individus) et balisée (par interventions), avant d'être convertie en données ASCII par l'intermédiaire du logiciel Lexica (*Sphinx Dev.*, vers. 5.0), lequel rassemble des programmes de lexicométrie, de statistique textuelle et d'analyse multivariée (Cf. Cibois, 2007 : 43 *sqq*, pour des explications plus générales). Le moment textuel reporté ci-dessous donne un extrait du corpus linéaire en données brutes avant que n'interviennent les opérations de tri et de conversion (« M » renvoie à l'adulte) :

> Maxim : qu'est-ce que tu fais Kyllian ?
> Kyllian : bah / qu'est-ce que je fais ? // paw paw paw
> Ambre : c'est laquelle la plus grande ?
> Maxim : bah / les grandes fusées ils vont décoller tous ensemble
> Ambre : c'est laquelle la plus grande ?
> M : où vous en êtes ? // ah je vois un petit Antony qui travaille très bien / qui a presque compris comment on pouvait faire // je suis très contente Anthony // maintenant regarde cette voiture / comment sont accrochées les roues ? // elles sont toutes du même côté ?
> Antony : non elles sont des deux côtés
> M : comment on peut faire pour accrocher les roues des deux côtés ? / est-ce que ces deux trous suffisent ? // combien Léo a fait de trous pour fabriquer sa voiture ?
> Kyllian : quatre
> M : quatre / et combien il ya de trous sur la voiture d'Antony ?
> Kyllian : deux
> M : où est-ce que je dois faire les trous qui manquent alors ?
> Anthony : ici

Comme il s'agissait principalement de dégager les domaines de variabilité interindividuelle entre les productions verbales des enfants, nous avons soustrait celles de l'adulte de manière à effectuer un premier dépouillement sur la base du nombre d'interventions (spontanées ou sollicitées), du nombre de mots produits et de la proportion de mots distincts présents dans les verbalisations collectées, lesquels dans l'ensemble représentent donc les premières variables prises en compte. À la faveur d'un premier traitement informatique, il est apparu que la plus ou moins grande dispersion des variables nominatives (renvoyant aux enfants), dans les groupes, révèle des rapprochements et des écarts significatifs

parmi les productions, mais aussi que les écarts plus ou moins prononcés entre les variables descriptives (interventions, mots et mots distincts) contribuent à dégager différents profils d'ateliers. Pour autant, un tel traitement s'est avéré en partie trivial, en ceci qu'il n'a fait que confirmer des résultats déjà pressentis, pour certains, par les intervenants, tout en en donnant toutefois des garanties numériques irréfutables.

Il convenait, de ce fait, de recourir à une analyse linguistique plus aboutie, de sorte, notamment, à mesurer la trivialité du premier traitement, mais aussi à distinguer ce qui, parmi les données, distingue les productions de ce qui renvoie spécifiquement à la structuration discursive. Les variables descriptives retenues ont alors été les verbes (composés ou non, négativés ou non), les éléments proprement thématiques et rhématiques, les mots ou les groupes de mots « encadrant » le discours, et enfin les opérateurs tels que les coordonnants ou les subordonnants. La caractérisation de ces éléments sur les données brutes non linéaires nous a conduit à délaisser pour un temps les annotations, les multitranscriptions et les commentaires para-verbaux, lesquels ont été pris en compte au moment du tri (Torterat, 2010a). Par commodité, nous avons « indiqué » les items avec diverses représentations graphiques : les verbes et « amas verbaux » le sont en italiques, les éléments thématiques en majuscules, rhématiques en petites majuscules, les « cadratifs » en souligné, et les opérateurs spécifiques en gras, ce qu'exemplifie l'extrait ci-dessous (productions verbales de « Lucas », dans le « groupe 2 ») :

c'est CASSE
c'est DROLE
il y a ~~euh~~ // UNE BOITE ?
JE *prends* CELLE-LA
JE *prends* LES BRINDILLES | A LA MER **et** LA / | LES BRINDILLES | A LA MONTAGNE
ON *dirait* UNE EPEE // ~~tayaya~~
OUI *c'est* VRAI ÇA *marche*
OUI
regarde **comment** ELLE *est* ma voiture
VOITURES
ON *joue* A LA COURTE PAILLE ?

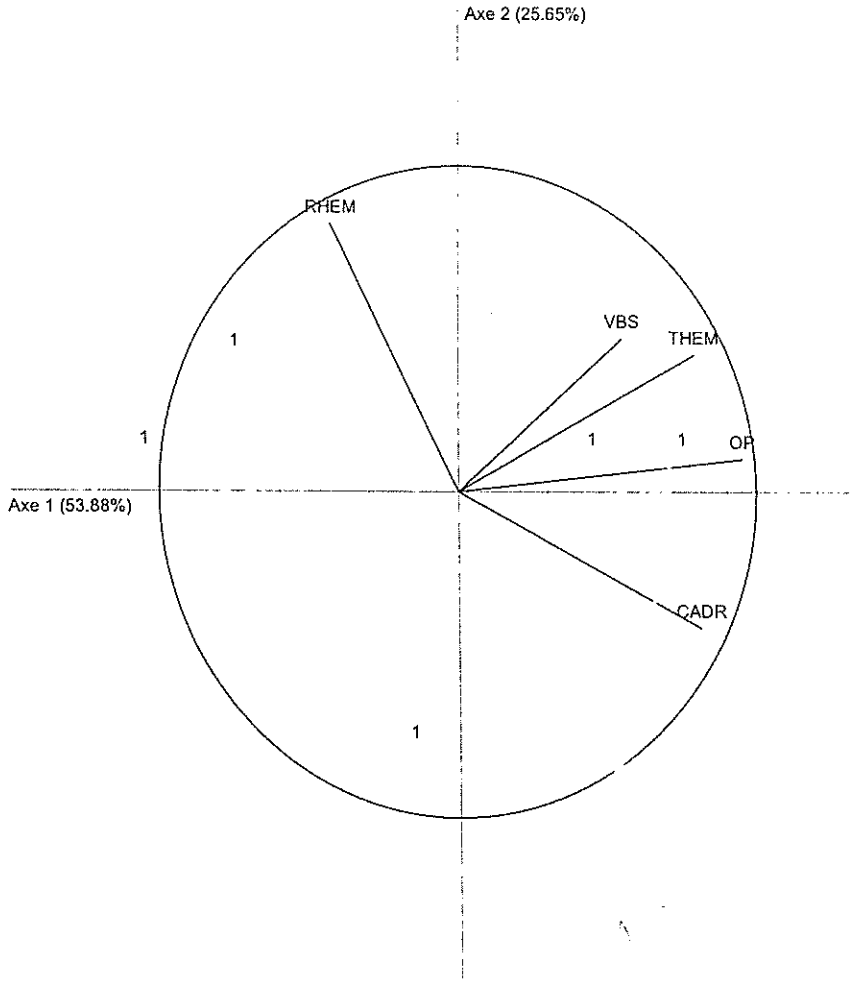
Ces caractérisations posent un certain nombre de difficultés, dans la mesure en particulier où 3 à 6 pour cent des éléments peuvent être, malgré les informations contextuelles, décrites de plusieurs manières. Une fois quantifiées, les proportions par enfant donnent toutefois les répartitions suivantes (pour le « groupe 2 ») :

	VBS	THEM	RHEM	CADR	OP
ALAIN	8,00%	12,00%	4,00%	0,00%	0,00%
AMBRE	31,00%	28,00%	23,00%	41,00%	34,00%
ANTONY	1,00%	2,00%	6,00%	0,00%	0,00%
KYLLIAN	16,00%	16,00%	10,00%	18,00%	12,00%
LUCAS	15,00%	12,00%	15,00%	6,00%	12,00%
MAXIM	21,00%	24,00%	25,00%	29,00%	36,00%
OPHELIO	8,00%	6,00%	17,00%	6,00%	6,00%

On remarque ainsi que Lucas, par exemple, emploie 15 pour cent des verbes et des éléments rhématiques produits par le groupe, 12 pour cent des éléments thématiques et des opérateurs, mais seulement 6 pour cent des cadratifs. Certes, nous ne reportons là qu'un « instantané » mais celui-ci, confirmé en grande partie par les deux autres, témoigne de possibles acquisitions et positionne les productions verbales de l'enfant dans un cadre dont on peut déjà donner quelques conclusions intermédiaires.

La démarche analytique mise en œuvre a donc été « multivariée », en ceci qu'elle a impliqué des composantes multiples, comme c'est le cas en sociologie par exemple, où elle permet de confirmer l'existence d'habitudes culturelles, de groupes sociaux ou de comportements récurrents, parmi lesquels se dégagent éventuellement des tendances ou des regroupements non prévisibles *a priori*. Au moment d'en fournir une version résumée, et vu que nous disposons d'éléments textuels convertis en données numériques, l'une des représentations graphiques possibles consiste dans ce que les statisticiens appellent une analyse en composantes principales (ACP).

Parmi d'autres avantages, l'ACP rend compte d'une multiplicité qu'elle unifie et qu'elle résume en quelques composantes, qu'elle « aplatit » en deux dimensions, quand bien même la plus ou moins grande « inclinaison » des vecteurs (ici nos variables descriptives) apporte quelques indications sur une autre dimension, elle-même plus ou moins confortée par la matrice des corrélations. L'ACP est par ailleurs factorielle quand elle passe par le report de facteurs résumant, pour leur part, les différentes valeurs intégrées dans les données. Les pourcentages des facteurs nous informent de ce fait dans quelles proportions ces derniers « expliquent » les variances existant entre les valeurs. Ainsi, plus la somme des pourcentages approche de 100, plus il apparaît que les facteurs expliquent l'ensemble des données, ce dont la représentation suivante, qui renvoie aux données précédemment documentées, donne un bref aperçu :



Prise comme telle, cette carte témoigne du fait que les facteurs expliquent les taux de variance de manière satisfaisante, vu que leur somme dépasse les 79 pour cent. Elle permet aussi de représenter comment se positionnent les variables descriptives les unes par rapport aux autres, tout en apportant quelques informations sur les écarts existant entre les variables nominatives (ici les *observations* en 1).

Cela étant, l'apport de cette ACP demeure aporétique si elle n'est pas complétée par d'autres quantifications. Ainsi, le pourcentage d'explication de la variance par les facteurs n'apporte aucune information définitive s'il n'est pas confronté à ce qu'on appelle les « contributions » (*positives* et *négatives*) qui leur sont assignées. Pour reprendre notre exemple, le premier axe reçoit ici les

contributions positives des opérateurs (33 %) et des cadratifs (24 %), et une contribution négative des éléments rhématiques (- 6 %), alors que le deuxième reçoit surtout les contributions des éléments rhématiques (52 %) et verbaux (17 %), et négativement celle des cadratifs (- 14 %). Toute déduction qui ne tiendrait pas compte de ces informations peut mettre à mal les déductions qu'en retire l'analyste. Or, d'autres données statistiques, comme la variance expliquée par toutes les composantes (qui sont résumées dans la carte), ainsi que les pourcentages cumulés et les corrélations, sont tout aussi significatives.

L'ensemble de ces apports sont donc autant de possibilités pratiques, qui doivent néanmoins être combinées pour garantir l'analyse et dépasser les difficultés de tri, de classement et de quantification qui se présentent à l'analyste. Dans le cas présent, il s'est avéré que non seulement les deux traitements opérés, sans se contredire, méritent d'être combinés, mais encore qu'ils nécessitent de constants retours vers les éléments de corpus, et donc en partie les données brutes. Qui plus est, l'analyse multivariée pratiquée sur de tels corpus discursifs demeure tout à fait incomplète si elle n'est pas confortée par des données longitudinales (autrement dit des combinaisons d'instantanés collectés à des moments suffisamment distants), à défaut de quoi les productions enregistrées ne permettent pas de conclure sur d'éventuelles acquisitions.

3.2. Deuxième exemple : des entretiens d'animateurs en sociologie des professions

Le deuxième type de corpus discursif que nous présentons ci-après renvoie à des entretiens menés auprès d'animateurs de l'environnement. Ces entretiens, établis en partenariat avec des sociologues, des linguistes et des chercheurs en sciences de l'éducation, mais aussi des représentants d'associations diverses, ont pour objet de dégager les contours d'une profession encore fragile et peu représentée dans les pays européens. Concrètement, les animateurs concernés interviennent sur le terrain pour intéresser les enfants et les adolescents à des questions de vie en société, de respect de l'environnement et pour partie d'urbanisme. À ce titre, ils organisent des animations pédagogiques, comme des visites ou des interventions en classe d'élèves, mais aussi des débats intermunicipaux, de manière à participer à la sensibilisation de publics variés. Dans cette vue, l'Agence Nationale de la Recherche (ANR), qui, sur le territoire français, est chargée, parmi d'autres mandats, de soutenir la recherche et ce qu'on appelle un peu commodément son *utilité publique*, parraine, dans le même temps, une analyse historico-critique des discours portant sur ces mêmes domaines de recherche. Pour ce qui concerne ici la question sociologique, il s'agit notamment de porter un regard critique sur l'écocitoyenneté, le « développement durable » et leur enseignement, ce qu'avait permis auparavant

l'ADEME, dont nous avons parlé. Les équipes constituées ont toutes, dans ce cadre, des mandats distincts : en plus de groupes qui se rassemblent sur la méthodologie et l'épistémologie des thématiques, d'autres réfléchissent à l'organisation pratique des « débats citoyens », ou encore sur les parcours individuels des intervenants.

Effectivement, la profession d'animateur de l'environnement fait partie de ces métiers « émergents » qui, comme certaines interventions sociales ou certaines professions en lien avec l'« aide à la personne » (De Robertis 2007, Barbe *et al.* 2008, Ion 2009), rejoignent des préoccupations sociétales de plus en plus partagées dans les pays industrialisés. Or cette profession est, dans le même temps, en voie de reconnaissance. Les personnes sollicitées ressentent donc quelque embarras à s'expliquer à la fois sur les parcours personnels qui sont les leurs et sur leurs aboutissements professionnels, et se présentent plus volontiers comme les représentants d'une culture que d'une profession à proprement parler. On le remarque assez facilement dans les discours qu'ils tiennent sur ces questions, d'autant que leurs réponses, la plupart du temps, se révèlent simultanément explicatives et digressives, en ceci que les locuteurs opèrent de constants va et vient entre la description de leurs tâches quotidiennes et leur justification sociale (Torterat, 2010b). Cela étant, quand les animateurs sont rassemblés dans le cadre d'un même entretien, le dialogue qui s'établit témoigne à bien des égards d'une véritable co-construction de valeurs communes, avec des similitudes et des dissemblances entre les parcours individuels, les transitions professionnelles, les démarches pédagogiques qu'ils mettent en place au cours de leurs animations, mais également une absence quasi-générale d'énoncés parémiques ou axiologiques, laquelle distingue nettement les présents dialogues des discours que tiennent les intervenants sociaux sur leurs pratiques (Martinez, 2005 ; Martinez et Poydenot, 2007 : 59-62).

Concrètement, les entretiens sont retranscrits de la manière suivante :

199 MLM	*Vous êtes jeune, pour avoir fait tout cela... Excuse moi, tu as quel âge ?*
200 E	J'ai trente deux…
201 MLM	*C'est formidable... Tu as déjà fait tout ça ?*
202 E	Oui…
203 MLM	*Et maintenant donc, tu... Cela fait combien de temps que tu es ici ?*
204 E	Ça fait, ça fait, à peine plus d'un an, j'ai commencé en octobre 2005
205 MLM	*Tu es content (xx)*
206 E	Oui, c'est une manière, un petit peu de, de revenir dans ma région… Je suis originaire des Ardennes (xx)… J'avais du mal à concilier mes aspirations professionnelles un petit peu…(xx) contribuer quelque part au respect de l'environnement (xx)

Les interventions de l'intervieweur sont reportées en italiques, et malgré un enregistrement difficile (les *(xx)*, par convention, renvoient à des ensembles plurisyllabiques inintelligibles), la transcription respecte les suspensions, les ruptures de construction, les pseudo-tours (tels que *Mmm*), et donc quelques disfluences propres à l'oral. Ceux-ci pourraient paraître, à première vue, inopérants dans le domaine de la sociologie des professions, mais il n'en est rien : non seulement ces indications témoignent du fait que les aménagements pratiqués sur le corpus s'arrêtent à ménager des graphies correctes pour les mots, ainsi qu'une ponctuation propre à l'écrit tel qu'il est envisagé par le retranscripteur, mais aussi ne redresse ni les hésitations (qui peuvent révéler par exemple des réticences à nommer les éléments de réponse), ni les opérations qui ont une incidence directe sur la construction des énoncés (comme la topicalisation par exemple). Certaines des corrections effectuées laisseront éventuellement présumer qu'elles sont autant de réaménagements, pour reprendre un terme couramment employé, mais cela présente assez peu d'inconvénients pour la démarche analytique ici envisagée.

Effectivement, compte tenu du fait que le principal objectif de cette recherche consiste à saisir comment les locuteurs interviewés se représentent leur profession, en décrivent les contours, et se positionnent vis-à-vis de savoirs et de savoir faire auxquels ils s'identifient diversement, ce sont surtout les *moments* représentés dans le corpus interlocutif qui prennent sens. De ce point de vue, la matérialité des éléments s'appuie surtout sur les *tours de parole*, parmi lesquels on peut reprocher aux auteurs de ne pas avoir indiqué les cas de chevauchements. Il va sans dire que la représentativité du corpus, de son côté, est abondamment questionnée : la description fournie du terrain d'enquête et du contexte de la collecte, les explications apportées sur les conditions de production des données, sur les objets discursifs abordés, ainsi que sur les implications épistémologiques des lexèmes d'*environnement* et d'*écocitoyenneté*, occupent autant de place que les extraits du corpus de référence et l'analyse qui en est effectuée. D'autre part, l'un des enjeux de ce type de production en sociologie renvoie au regard qui peut être porté sur la conduite de l'entretien, à propos de laquelle la discussion reste ouverte, notamment quand l'intervieweur tente d'amener l'interviewé vers l'objet même de l'analyse :

258 E — Je ne sais pas si euh la fonction d'animateur à proprement parler me permet de chercher-comprendre… Par contre dans le CPIE (xx) … projets associatifs (xx) Disons que je suis rentré ici sachant que j'avais cette euh, cette ouverture éventuellement (xx)[4]

[4] Les CPIE sont des *Centres Permanents d'Initiatives pour l'Environnement*. Créés en 1973, ils forment un réseau de partenaires, en particulier des associations, qui sont regroupés par l'intermédiaire du label qui les identifie, et dont l'esprit se caractérise par

259 MLM	*C'est une autre carte, une autre mission en fait de spécialisation que tu aimerais explorer aussi euh et actuellement c'est une animation de la formation en sachant que tu y trouves des choses mais c'est pas, fondamentalement là que tu, que tu réaliseras peut-être les, tes aspirations. Et tu arrives à faire le lien pour l'instant entre ces deux... aspirations?*
260 E	Un petit peu...
261 MLM	*...de recherche ingénierie*
262 E	Comme je suis là depuis peu de temps
263 MLM	*...réflexion quand même sur les dispositifs, euh (xx)*
264 E	Là je prends un peu mes marques, hein, ça fait un an que suis là euh, c'est très intensif...(xx) des périodes (xx) J'ai pas encore eu le temps de m'inscrire dans tous les aspects du CPIE mais j'y ai quand même déjà participé, par exemple pour les contrats (xx) ... ou alors de choses plus précives / précises pardon, par exemple l'implantation de corps morts au fond de l'eau sur les sites de... plongée

Ce matériau nous permet d'assister à la co-construction d'un discours commun, auquel participent les interlocuteurs d'une réplique à l'autre. Il est ainsi facile de déduire de la conduite de l'entretien qu'elle s'apparente à bien des égards à une forme d'accompagnement, comme cela se pratique dans presque tous les cadres de professionnalisation. On notera néanmoins qu'avec une présentation qui s'appuie sur les variations à l'intérieur du corpus, les formes discursives se rassemblent autour d'objets spécifiques, au premier rang desquels intervient le *biographique*. Effectivement, les éléments pris en compte témoignent d'une biographisation du discours qui paraît tantôt commentative, tantôt narrative. Or, il s'agit là de deux démarches distinctes, que valorisent conjointement les interventions de l'intervieweur et celles de l'interviewé. Ainsi remarque-t-on que tantôt les participants répondent à la question de savoir ce que sont les contours de la profession, tantôt nous incitent à discerner dans quelle mesure s'établissent des liens plus ou moins directs entre leurs parcours personnels et les parcours professionnels qui sont les leurs :

191 F	Donc nous NOTRE ROLE EN TANT QUE PERMANENTS, en... C'est vrai que nous on est, nous on est quand même des PROFESSIONNELS...
192 MLM	*Mmm*

une transversalité des domaines, un décloisonnement des approches (ethnologie, écologie, urbanisme, *etc.*), et une (re)découverte permanente du « terrain vivant ».

193 F	Donc NOTRE ROLE C'EST AUSSI, DE FAIRE DE L'ACCOMPAGNEMENT DES ANIMATEURS. Si tu veux moi, je passe pas mal de temps avec eux, pendant tu vois, dans l'activité si tu veux pour moi, ils sont comme les gamins… (...)
199 F	Moi, c'est vrai que moi, par rapport à ça J'AI QUAND MEME L'EXPERIENCE DE LA VOILE, aussi, ça, ÇA SERT BEAUCOUP DANS LE CADRE DE L'ENCADREMENT puisque moi je m'occupais des… J'ETAIS MONITRICE DE VOILE, J'AI ETE ENTRAINEUR VOILE AUSSI, J'AVAIS UNE EQUIPE COMPETITION, EUH, CHEZ LES, EN PRE-ADOS
200 MLM	*Mm*
201 F	donc si tu veux CE COTE UN PEU, MANAGEMENT, COCOONING TU VOIS, D'UNE EQUIPE… C'est
202 MLM	(voix multiples) *entraîneuse de, enfin (xx)*
203 F	(xx) (voix multiples) Tu re-veux du café ?
204 MLM	*Euh, oui, merci… un petit peu…*
205 F	Tu vois bon c'est… Je pense que tout ce, TOUT CE VECU EN FAIT, MAIS ÇA AIDE ENORMEMENT ET APRES BAH ON VA TOUS ABORDER LES… NOTRE TRAVAIL DE MANIERE DIFFERENTE EN FONCTION DE CE QUE L'ON AURA FAIT AVANT, EN FONCTION DE, DE, DE PLEIN DE CHOSES.
206	xxx
207 F	Je pense que ça c'est… Ça tu devrais le… Ça tu dois le, le noter, enfin j'imagine assez souvent quoi… Que L'APPROCHE EN FAIT DES, DES INTERVENANTS ELLE EST DUE A TOUT CE QU'ILS AVAIENT VECU AVANT.

Si nous mettons de côté l'anecdote du café, et pour nous conformer à ce que nous avons présenté du corpus précédent, nous avons ci-dessus reporté en majuscules ce qui consiste dans une description de la profession, et en petites majuscules ce qui relève d'une biographisation de la réponse (avec un cas de recoupement possible en 205F). D'une part, cela conduit à estimer dans quelles proportions ces deux types d'éléments de réponse interviennent dans le corpus. D'autre part, ces répartitions permettent de donner une version concrète de la démarche interlocutive dans laquelle s'inscrivent les personnes interviewées.

Nous ne reporterons pas ici les conclusions auxquelles en sont venus les organisateurs des entretiens, qui ont notamment opéré plusieurs types de recoupements entre les données, certains par le biais d'analyses statistiques, d'autres à partir de commentaires, ou encore à partir d'une confrontation des présentes réponses à plusieurs documents d'archive. Dans tous les cas, la prise en compte des difficultés d'ordre méthodologique pour le traitement et l'analyse s'est avérée déterminante au moment de l'interprétation, qui a incité les

organisateurs à insister sur une nécessaire valorisation des personnes, ainsi que sur le caractère pressant d'une reconnaissance sociale, par l'ensemble des partenaires, d'une profession novatrice, mais en partie mésestimée.

En conclusion

Les corpus discursifs ont pour principal mérite de concerner presque tous les domaines de recherche en sciences de l'homme et de la société. Ils peuvent ainsi être sollicités pour démontrer comment certains locuteurs politisent un débat, narrativisent un fait quotidien, judiciarisent leurs propos, ou dans quelle mesure d'autres locuteurs, par exemple, ont acquis telle ou telle construction grammaticale ou telle ou telle capacité pédagogique. Dans cette vue et comme nous l'avons suggéré, ils sont à bien des égards incontournables.

À ce titre, rappelons, en marge des réflexions qui précèdent, qu'il convient de préciser à quoi servent concrètement les matériaux retenus. Dans le cas où le traitement et l'analyse de ces derniers aboutit à des conclusions qui leur sont propres, ils sont envisagés intégralement et, suivant les domaines de recherche, font l'objet d'explications plus ou moins abondantes sur leur représentativité et sur leur matérialité. Dans celui où ils contribuent à la confrontation d'archives, ils prennent place dans un historique, mais aussi figurent parmi d'autres corpus qui ne renvoient peut-être pas complètement aux mêmes contextes, ou ne répondent pas forcément aux mêmes conditions de collecte, ce qui rend la confrontation, tout traitement mis à part, particulièrement onéreuse en termes de commentaires et d'éventuels ajustements. Une autre possibilité, qui apparaît couramment dans certains domaines, est celui où les corpus discursifs servent d'exemplification. Dans ce cas, ils sont non intégraux et représentent des appoints, mais suscitent généralement peu d'explications sur ce qui précède leur production.

Bien d'autres possibilités existent, ce qui rend les recours à ces supports aussi opportuns que variés. Parmi les questions que cela pose, l'une d'elles consiste à définir ce qui en légitime l'emploi à proprement parler. Car si les éléments de corpus en profitent généralement très concrètement, ils ne sont pas faits pour valoriser les traitements qu'on en effectue. La démarche analytique aura ainsi d'autant plus de mérite qu'elle permettra de se saisir de ce qui fonde les singularités des productions.

Bibliographie

Avanzi, Mathieu et Anne-Sylvie Horlacher (éds). (2007). *Structuration grammaticale et structuration discursive,* Numéro thématique de *Tranel* 47. Université de Neuchâtel.

Barbe, Laurent *et al.* (2008). *Intervenants sociaux et analyses de pratiques.* Paris : L'Harmattan.

Baude, Olivier (éd.). (2006). *Corpus oraux. Guide des bonnes pratiques 2006.* Paris : CNRS éditions.

Blanche-Benveniste, Claire *et al.* (1990). *Le Français parlé : Études grammaticales.* Sciences du langage. Paris : CNRS Editions.

Branca-Rosoff, Sonia. (1998). *Le Mot, analyse de discours et sciences sociales.* Aix : PUP *Langues et Langage* 7.

Bulot, Thierry et Vincent Veschambre (éds). (2000). *Mots, traces et marques. Dimensions spatiale et linguistique de la mémoire urbaine.* Paris : L'Harmattan.

Cibois, Philippe. (2007). *Les Méthodes d'analyse d'enquête.* Paris : PUF.

De Robertis, Cristina (éd.). (2007). *Méthodologie de l'intervention en travail social : l'aide à la personne.* Paris : Bayard.

Dister, Anne et Anne-Catherine Simon. (2008). « La Transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé ». *Arena Romanistica* 1/1 : 54-79.

Gadet, Françoise. (2003). *La Variation sociale en français.* Paris : Ophrys.

Grobet, Anne. (1997). « La Ponctuation prosodique dans les dimensions périodique et informationnelle du discours ». *Cahiers de Linguistique française* 19 : 83-123.

Ion, Jacques, (2009). « Travailleurs sociaux, Intervenants sociaux : quelle identité de métier ? ». *Informations sociales* 152-2 : 136-142.

Jackiewicz, Agata. (2002). « Repérage et délimitation des cadres organisationnels pour la segmentation automatique des textes ». *CIFT'02* : 95-107.

Longhi, Julien. (2007). « L'Objet discursif *intermittent* : construction d'une forme sémantique et évolution des topoï dans un corpus de presse ». In G. Cislaru, O. Guérin, K. Morim, E. Nee, T. Pagnier et M. Veniard (éds). *L'Acte de nommer – une dynamique entre langue et discours.* Paris : PUPS, 149-163.

Mann, William et Sandra Thompson. (1988). « Rhetorical Structure Theory : towards a functional theory of text organisation ». *Text* 8 : 243-281.

Martinez, Marie-Louise. (2005). « Le Débat comme espace interlocutif d'identification des textes et des personnes ». *TRÉMA* 24 : 77-101.

Martinez, Marie-Louise et Frédéric Poydenot. (2007). « Accompagner les identités, écocitoyenneté et santé ». *comm. au Congrès de L'AREF*, Strasbourg.

Pêcheux, Michel. (1990) : *L'Inquiétude du discours.* Paris : éditions des Cendres (textes sélectionnés et présentés par Denise Maldidier).

Pincemin Bénédicte. (2007). « Introduction » au numéro 6 de la revue *Corpus* : 5-15 [article consulté le 6 juillet 2009] : http://corpus.revues.org/index812.html .

Rastier, François. (2007). « Passages », *Corpus* 6 : 125-152 (consulté le 3 octobre 2008) : http://corpus.revues.org/index832.html .

Roulet, Eddy. (2002). « Le Problème de la définition des unités à la frontière entre le syntaxique et le textuel ». In Actes du Congrès *Y a-t-il une syntaxe au-delà de la phrase ?* (Paris, 09/2000) : 161-178.

Shriberg, Elizabeth *et al.* (2000). « Prosody-Based Automatic Segmentation of Speech into Sentences and Topics ». *Speech Communication* 32-1 : 127-154.

Sinclair, John. (2005). « Meaning in the Framework of Corpus Linguistics ». In W. Teubert (éd.), *Lexicographica*, Tubingen, Niemeyer : 20-32.

Torterat, Frédéric. (2009). « La *Dictée à l'adulte*, telle que pratiquée à l'École : une approche combinée des faits grammaticaux et des phénomènes prosodiques ». *Travaux linguistiques du Cerlico* 22, Presses Universitaires de Rennes : 293-308.

Torterat, Frédéric. (2010a). « Une Analyse multivariée des productions verbales de jeunes enfants : un élément de plus en faveur des corpus longitudinaux ». In G. Williams (éd.), *Actes des 6èmes Journées de Linguistique de Corpus*. Lorient : Université du Sud-Bretagne.

Torterat, Frédéric. (2010b). « Les Récits de vie des animateurs de l'environnement : des parcours personnels aux parcours professionnels ». In M.L. Martinez et J. Girault (éds.). *Les Métiers de l'environnement*. Paris : L'Harmattan.