

L'ANALYSE MULTIVARIÉE DES PRODUCTIONS VERBALES DE JEUNES ENFANTS : UN ÉLÉMENT DE PLUS EN FAVEUR DES CORPUS LONGITUDINAUX

Frédéric Torterat
Université de Nice / IUFM
frederic.torterat@unice.fr

1 INTRODUCTION

Les verbalisations des jeunes enfants, en particulier dans un cadre dialogal, témoignent d'acquisitions diverses. La prise en compte de cette diversité, par les personnes qui encadrent les enfants et qui les accompagnent, suppose d'en retenir plusieurs domaines de variabilité, laquelle est interindividuelle en ceci que les acquisitions ne sont pas les mêmes d'un enfant à l'autre, mais aussi intra-individuelle, en ceci que les capacités verbales d'un même enfant sont l'objet de constantes modifications. A cet égard, si l'analyse du premier domaine de variabilité peut s'appuyer sur ce que nous nommerons des corpus instantanés, celle du deuxième implique la plupart du temps des corpus longitudinaux (plusieurs instantanés d'un moment M1 à un moment Mn, et pour les mêmes enfants).

D'une manière assez générale, l'approche de ces types de corpus ne s'avère productive que dans la mesure où elle aboutit à des éléments révélateurs, que nous appellerons, comme cela se pratique par ailleurs, significatifs en acquisition. Quand ils sont convertis en données proportionnelles, les éléments discursifs appellent plusieurs formes de traitement, dont il est quelquefois difficile de prévoir les profits pour la pédagogie ou la recherche (Sinclair 2005, Pincemin 2007). Dans cette vue, l'une des possibilités qu'apporte aujourd'hui le traitement informatique des données, à savoir, pour ce qui nous occupe ici, l'analyse multivariée, permet d'établir des liaisons statistiques entre plusieurs variables, et de rendre compte, quand la démarche présente un minimum de garanties méthodologiques, d'un système de relations à l'appui desquelles on peut envisager de désigner des caractéristiques communes d'une part, et d'autre part distinctes, autrement dit, plus simplement, des rapprochements et des écarts.

Couramment employé en sociologie, et de plus en plus en linguistique de l'acquisition, ce type de traitement, qui est *multivarié* en ceci qu'il implique des composantes multiples, confirme par exemple l'existence de groupes sociaux, d'habitudes culturelles ou de comportements interindividuels parmi lesquels se dégagent éventuellement des tendances (Cibois 2007). Il en est ainsi de pratiques qui présentent simultanément les mêmes caractéristiques, ou de groupes d'individus qui ont un comportement analogue : autant d'éléments qu'une représentation unifiée permet d'aborder dans un ensemble où s'organisent des liaisons plus ou moins effectives que les statisticiens, à la suite des travaux de différents mathématiciens (voir Martin 1997), désignent à travers des attractions et des répulsions entre un nombre plus ou moins élevé de variables.

Rappelons toutefois que les résultats d'une analyse multivariée (désormais AM) ne répondent à « aucune métrique simple », pour reprendre l'expression de Rastier (2009), et qu'ils apportent peu d'informations s'ils ne rejoignent pas assidûment les données du corpus, et donc le corpus lui-même. Qui plus est, pour peu que les variables renvoient à des données

brutes ayant des liens discutables avec l'objet de l'analyse, un tel traitement informatique a toutes les chances de présenter un certain effritement. Enfin, dans le cas où des regroupements s'organisent entre des variables renvoyant à des allants-de-soi particulièrement improductifs, les conclusions intermédiaires peuvent déboucher sur des facteurs triviaux, qui sont d'un gain minime.

D'autres défauts peuvent apparaître au cours du traitement, et c'est donc suivant la minutie avec laquelle s'effectue l'approche méthodologique des données brutes que ces deux objets de mésestime – dispersion de la recherche et trivialité des composantes représentées – seront pris en compte ou non.

Car les atouts d'une AM ne sont pas négligeables : d'une part, ce type d'analyse décrit une multiplicité à partir de laquelle elle dégage des regroupements ; d'autre part, elle conduit à concrétiser, à partir d'un corpus converti par exemple en plusieurs (sous-)ensembles de proportions, des tendances quantifiables à travers ce qu'on appelle la « contribution aux facteurs », avec des domaines de variance différents selon les combinaisons de variables opérées.

Le corpus présenté dans ces pages n'est pas extrait d'une base de données généraliste de type CHILDES (Mac Whinney 2000) ou de données précédemment converties, mais porte sur les productions spontanées d'enfants enregistrés dans un contexte interlocutif de terrain, de manière à fournir des éléments de corpus oraux sur des bases accessibles pour des non statisticiens, et ce dans un temps court. Cette approche permet aux informations d'être employées spécifiquement dans le cadre de la professionnalisation des professeurs des écoles, d'autant que l'annotation des corpus oraux n'est pas sans poser des problèmes spécifiques (Benzitoun 2004), et que la multi-annotation exige une autre forme d'accompagnement. Collecté auprès de classes de maternelle à Nice (en circonscription), le corpus en question, intitulé DAM07, est constitué de productions correspondant à trois « dialogues de classes » encadrés, dans les mêmes demi-journées et sur des supports pédagogiques analogues, par M (professeur des écoles à Nice en 2007). Les effectifs « nominatifs » renvoient donc à des groupes restreints de petites et moyennes sections de maternelle, avec dans notre cas des enfants âgés de 2,5 à 4,2 ans, et des matériaux qui représentent un « petit » ensemble (expérimental) de 12000 mots.

Qu'on nous permette d'indiquer que les objectifs professionnels de cette recherche ont d'abord été praxéologiques : il s'agissait de déterminer comment organiser les groupes de manière à prendre en compte la diversité des acquisitions verbales des enfants, avec comme possibilité de mesurer les (ré)emplois des marqueurs de structuration discursive à travers l'analyse de cette diversité. Les données – hors caractéristiques phonologiques – ont donc suscité plusieurs démarches, dont nous reportons ci-après le suivi et les premiers résultats.

2 APPROCHE MÉTHODOLOGIQUE

Le corpus discursif sur lequel nous nous appuyons renvoie à des productions verbales qui, au moment des retranscriptions, ont été faiblement redressées. Etant donné que les données acoustiques et prosodiques n'ont pas été prises en compte, la ponctuation qui a été employée dans le format « texte », pour ce qui la concerne, n'apporte de son côté aucun recours explicatif ou descriptif. Pour autant, les éléments collectés nous ont parus suffisamment robustes pour l'analyse : sans doute n'est-ce pas la « grandeur » du corpus qui garantit sa représentativité (à moins de ne l'envisager qu'à travers les chaînes de caractères), mais sa consistance pour l'analyse. Pour celle-ci, nous avons envisagé la diversité des productions discursives bien sûr, mais aussi les variables qu'elles sont en mesure d'impliquer, les regroupements qu'elles permettent d'envisager, ainsi que leur apport d'informations non

préétablies. Cette démarche, dans l'ensemble, s'est alors assigné pour objets de dégager des données quantitatives qui, dans un deuxième temps, ont été confrontées à des éléments caractérisés de manière plus aboutie.

En voici tout de suite un extrait pour illustration, lequel est reporté en « texte linéaire » tel qu'il a été retranscrit une première fois, sans les annotations ni les remplacements correctifs (l'abréviation « M » renvoie à l'intervenante) :

G1 : Les coqs (enfants grands à moyens parleurs). Sont présents : Marie-Sarah, Léo, Noémie, Mattéo B, Apolline, Mathilde.

Léo : Qu'est-ce que tu vas enregistrer ?

M : Je vais vous enregistrer pendant que vous fabriquez les voitures. Alors, dites-moi, vous avez trié ce qui roulait ?

Noémie : Moi c'est dans la boîte !

Marie-Sarah : Moi aussi c'est dans la boîte.

M : Pour fabriquer une voiture, qu'est-ce qu'on doit prendre ?

Mattéo : Des roues !

M : On doit prendre des roues...

Léo : Et aussi la carrosserie.

M : La carrosserie aussi, oui tu te souviens bien Léo. Alors choisissez tous une boîte pour faire la carrosserie de votre voiture.

Mathilde : Moi je prends cette boîte.

Noémie : Et moi cette boîte.

M : D'accord. Comment vous allez vous y prendre pour accrocher les roues maintenant ?

Noémie : Ben on va faire un petit trou. (...)

Les aménagements graphiques que nous avons pratiqués sur le corpus linéaire sont restés sommaires, d'autant que le moindre ajustement réclamant une justification (Dister et Simon 2008), nous avons tâché de ne pas compromettre le recours aux documents par des explications trop abondantes. D'autre part, nous avons reporté dans les premières transcriptions une ponctuation conforme à une présentation « académique » des verbatim, telle qu'elle apparaît dans de nombreuses monographies, compte tenu du fait, notamment, qu'il n'existe pas de correspondances régulières entre les phénomènes prosodiques, comme les contours intonatifs et les pauses, et la ponctuation graphique (Grobet 1997, Shriberg *et al.* 2000).

Dans la mesure, d'autre part, où les enregistrements ont été effectués dans de bonnes conditions matérielles, nous avons mis momentanément à l'écart de cette première approche les multitranscriptions, les disfluences répétitives, les commentaires para-verbaux et les cas de chevauchements. Sur le plan méthodologique, cela revient à dire que la réflexion menée a moins porté sur le traitement d'un corpus oral spécifique, que sur la mise en œuvre d'une démarche analytique.

Les productions verbales ont été intégrées à la base de données du logiciel Lexica (5.0), diffusé par *Sphinx Développement*, et acquis à la suite d'un appel d'offre de l'Université de Nice à travers sa composante en charge de la formation des enseignants. Comme d'autres logiciels de même type, Lexica rassemble des programmes de lexicométrie, de statistique textuelle et d'analyse multivariée qui facilitent en partie la détermination de normes de dépouillement déjà pratiquées par ailleurs en linguistique de l'acquisition (Cf. Sansonetti 2003). Etant donné que la conduite méthodologique engagée, dans ses grandes lignes, a d'abord consisté à dégager les domaines de variabilité interindividuelle entre les enfants, nous

avons soustrait M non pas du corpus linéaire, mais des éléments triés en données brutes. C'est donc à partir du corpus non linéaire (trié par ordre alphabétique balisé) et ouvert sous la forme d'un fichier de données « L » (*texte à analyser* sous Lexica), que le format texte précédemment balisé est passé en données ASCII.

Dans l'application pratique, au moment de confirmer le choix automatique des variables (les intervenants d'un côté, et les données textuelles de l'autre), on requiert des options maximales pour un « texte », en paramétrant peu à peu les balises⁶⁷. Les variables apparaissent alors en « ouverte texte » (une *modalité* qu'il est possible de modifier par la suite). Une fois un premier dépouillement effectué sur l'instantané qui nous a servi de référence, les variables donnent les répartitions suivantes, lesquelles en représentent un premier abord.

Pour le Groupe 1 :

---	Nombre de mots	Nombre moyen de mots	Nombre de mots différents	Nombre de mots uniques	Fréquence maximum	Mot le plus fréquent
Apolline	40	4,44	12	11	2	roues
Léo	99	4,30	34	25	4	faut
Marie-Sarah	51	4,25	12	7	5	deux
Mathilde	51	4,25	21	18	2	roule
Mattéo	121	5,26	34	19	5	faire
Noémie	158	5,64	43	31	4	roule ⁶⁸

Pour le Groupe 2 :

---	Nombre de mots	Nombre moyen de mots	Nombre de mots différents	Nombre de mots uniques	Fréquence maximum	Mot le plus fréquent
Alain	22	3,67	10	6	2	Gaad
Ambre	105	4,20	31	20	4	Euh
Antony	12	3,00	6	5	2	deux
Kyllian	54	3,86	21	16	4	ça
Lucas	55	5,00	21	18	2	prends
Maxim	98	5,16	29	21	4	roues
Ophélio	39	2,79	18	16	5	roues ⁶⁹

⁶⁷ Si ce n'est : « l'ordre des balises sépare les observations » / « les balises sont numérotées et des parties créées ».

⁶⁸ Mots les plus fréquents :

Apolline : roues (2) ; petite (1) ; roule (1) ; arrive (1) ; veux (1) ; sais (1) ; faire (1) ; oui (1) ; grande (1) ; baguette (1) ;

Léo : faut (4) ; quatre (3) ; faire (3) ; regarde (3) ; ouais (2) ; roule (2) ; voiture (2) ; Julie (2) ; oui (2) ; Ah (1) ;

Marie-Sarah : deux (5) ; ça (2) ; veux (2) ; besoin (2) ; roues (2) ; Ah (1) ; oui (1) ; belle (1) ; mets (1) ; ici (1) ;

Mathilde : roule (2) ; boîte (2) ; oui (2) ; ça (1) ; débloque (1) ; ici (1) ; sais (1) ; biscuits (1) ; roue (1) ; marche (1) ;

Mattéo : faire (5) ; roues (4) ; ça (3) ; tchou (2) ; fait (2) ; vais (2) ; mettre (2) ; quatre (2) ; trou (2) ; arrive (2) ;

Noémie : roule (4) ; faire (4) ; ben (3) ; ça (3) ; non (3) ; oui (3) ; trou (2) ; boîte (2) ; roues (2) ; réussi (2)

⁶⁹ Mots les plus fréquents :

Alain : Gaad (2) ; fait (2) ; trou (2) ; veux (2) ; ça (1) ; roule (1) ; peux (1) ; aider (1) ; roue (1) ; faire (1) ;

Ambre : Euh (4) ; maîtresse (3) ; trous (3) ; roule (3) ; roues (2) ; fait (2) ; grande (2) ; non (2) ; veux (2) ;

Pour le Groupe 3 :

---	Nombre de mots	Nombre moyen de mots	Nombre de mots différents	Nombre de mots uniques	Fréquence maximum	Mot le plus fréquent
Alain	28	2,33	11	7	3	Heu
Jaed	11	1,83	9	8	2	Ah
Mathis	8	2,67	5	4	2	veux
Sarah	23	2,88	11	7	3	oui
Shon	7	1,00	5	3	2	Ata ⁷⁰

Ces données ne sont pas, bien entendu, à mettre sur un même plan. Si les deux premières variables renvoient à des données strictement quantitatives, la troisième coïncide indirectement avec la diversification des productions, et la suivante avec celle de possibles acquisitions lexicales. Les résultats 4 à 6 représentent pour leur part un apport difficile à mesurer, d'autant que, pour ce qui relève du nombre de mots uniques par exemple, d'éventuelles faiblesses quantitatives peuvent en revanche témoigner, chez l'enfant, d'une certaine démarche interlocutive. Les trois derniers types d'informations (*mots uniques*, *fréquence*, *mot le plus fréquent*) sont par ailleurs directement liées aux exigences pédagogiques à l'oeuvre au moment du relevé, et varient suivant la pratique évaluative éventuellement mise en place, les consignes de l'exercice et les conditions de leur passation. D'autre part, à la faveur d'une première confrontation des données, il est apparu que le « nombre moyen de mots » (par intervention) n'apporte en général qu'une corroboration quantitative des données qui l'encadrent. Ces éléments ont donc été, au moins pour un temps, considérés comme intermédiaires. Il en a été de même pour les répartitions par *catégories*, qui, sur corpus en données brutes, spécifient la liste des 10 mots marqués du lexique (avec les nombres d'occurrences pour chaque mot), ici pour *Apolline* :

pas	4	ai	1	arrive	1	est	1
faire	1	peuvent	1	roule	1	rouler	1
sais	1	veux	1				

regarde (2) ;
 Antony : deux (2) ; bâtons (1) ; ici (1) ; roues (1) ; non (1) ; côtés (1) ;
 Kyllian : ça (4) ; marche (4) ; Regarde (4) ; Paw (2) ; fait (2) ; beau (1) ; Bah (1) ; fais (1) ; bâtons (1) ; deux (1) ;

Lucas : prends (2) ; brindilles (2) ; oui (2) ; cassé (1) ; drôle (1) ; euh (1) ; boîte (1) ; mer (1) ; montagne (1) ;
 dirait (1) ;
 Maxim : roues (4) ; trous (4) ; faire (3) ; fais (3) ; fait (2) ; Kyllian (2) ; oui (2) ; Quatre (2) ; ça (1) ; Bah (1) ;
 Ophélio : roues (5) ; oui (2) ; accrocher (1) ; ça (1) ; deux (1) ; après (1) ; mets (1) ; Heu (1) ; Quatre (1) ;
 manque (1)

⁷⁰ Mots les plus fréquents :

Alain : Heu (3) ; Julie (3) ; ça (2) ; roule (2) ; Aye (1) ; mal (1) ; veux (1) ; arrive (1) ; regarde (1) ; roue (1) ;
 Jaed : Ah (2) ; oui (1) ; liou (1) ; veux (1) ; ça (1) ; mou (1) ; Yo (1) ; ya (1) ; bou (1) ;
 Mathis : veux (2) ; maison (1) ; oui (1) ; pique (1) ; Papa (1) ;
 Sarah : oui (3) ; deux (2) ; faut (2) ; Julie (2) ; belle (1) ; euh (1) ; roues (1) ; ça (1) ; accord (1) ; regarde (1) ;
 Shon : Ata (2) ; Voiture (2) ; Heu (1) ; oui (1) ; roule (1)

elle est plus petite
 elle / elle roule
 et moi j'ai pas de roues
 j'y arrive pas
 moi aussi je veux
 moi je sais pas le faire
 oui celle là
 une grande baguette
 tes roues elles peuvent pas rouler ?⁷¹

Même si certains de ces éléments permettent d'envisager des regroupements, de manière à réordonner dans Lexica les données brutes en tableur, ils ne procurent qu'une visibilité réduite des domaines de variabilité qui nous occupent. Indiquons en revanche qu'un report des mêmes données concernant l'intervenante principale (« M »), montre de son côté dans quelles proportions l'adulte référent est effectivement intervenu parmi les verbalisations des enfants :

---	Nombre de mots	Nombre moyen de mots	Nombre de mots différents	Nombre de mots uniques	Fréquence maximum	Mot le plus fréquent
Groupe 1	928	13,45	178	112	15	roues
Groupe 2	889	12,01	149	88	22	faire
Groupe 3	610	14,52	139	99	9	ça

Ce dépouillement liminaire permet toutefois de noter les premières récurrences vers un « tableau à plat généralisé »⁷², qui, pour le groupe 1 (*Marie-Sarah ; Mathilde ; Mattéo ; Noémie ; Apolline ; Léo*), apporte des indications non négligeables, comme le font apparaître les extraits suivants :

⁷¹ Sur corpus lemmatisé, avec la liste des 9 mots marqués du lexique (et les nombres d'occurrences pour chaque mot), cela donne les données ci-après :

pas	4	rouler	2	arriver	1	avoir	1
être	1	faire	1	pouvoir	1	savoir	1
vouloir	1						

il être plus petit
 il / il rouler
 et moi j avoir pas de roue
 j y arriver pas
 moi aussi je vouloir
 moi je savoir pas le faire
 oui celui là
 un grand baguette
 ton roue il pouvoir pas rouler

⁷²La manipulation informatique revient à se rendre à l'onglet « analyser », puis à cliquer sur « déterminer l'analyse » en indiquant les premières variables (ensemble des participants excepté M). Nous remercions chaleureusement Jean-Jacques Legendre pour ses explications sur tous ces points.

verbalisations
et aussi la carrosserie
et bien moi elle roule ma voiture
et c'est quoi ça / cet outil // c'est comme un bouchon
et moi aussi là / partout
et moi cette boîte
et moi j'ai pas de roues
moi aussi c'est dans la boîte
moi aussi je veux
moi ça me fait même pas mal
moi ça y est
moi c'est dans la boîte
moi c'est la boîte d'oeufs // ça c'est une boîte d'oeufs
moi ici
moi je l'ai fait
moi je peux pas / regarde
moi je prends cette boîte
moi je sais pas le faire
moi je vais les faire
moi je veux la mienne là
moi j'ai réussi à le faire
moi j'ai réussi à les faire
moi j'ai une roue qui marche pas très bien
moi j'arrive pas à le faire
moi tu m'as donné un grand // fais voir ton bâton
moi / là / là

En marge du caractère hasardeux et accumulatif de ces reports de productions non nominatives, ces énumérations nous conduisent à admettre la difficulté d'opérer des regroupements à partir de certains items. En outre, ces supports dénoncent le fait que les caractéristiques paradigmatiques, mais aussi syntagmatiques et topologiques, peuvent toutes avoir une capacité classificatoire. Cela suppose que nous écartions un relevé des « segments répétés » au profit d'un relevé d'éléments « récurrents », pour le traitement informatique desquels une identification manuelle s'impose.

Un tel tableau à plat généralisé valorise en partie, dans le même temps, la diversité des éléments du corpus, d'une part, et une possible représentativité des productions verbales, à partir desquelles il est intéressant d'établir des concordances spécifiques de manière à saisir la singularité de certaines données. Ainsi, des récurrences apparaissent quand l'opérateur interphrastique de transition *et (moi / aussi)* marque un lien paratactique entre le cotexte antérieur et les éléments qui lui sont postposés : tantôt thématissant (*et bien moi / et c'est quoi ça / et moi aussi là*), tantôt thématissant et rhématisant (*et moi cette boîte*), il présente dans plusieurs cas une dimension cadrative, en ceci qu'il contribue à la structuration discursive du dialogue et des co-interventions des participants (Saussure et Sthioul 2002 *inter al.*). De même, quelques récurrences interviennent à l'appui du pronom *moi*, dont les multiples

combinaisons avec des déictiques (*ici, là, c'est..*) témoignent variablement de la prise en compte effective, par les enfants, du contexte interpersonnel. Ce monosyllabe est d'ailleurs suivi dans la plupart des cas d'un pronom dialogal (*je, je (là), me, tu*), et suppose dans presque tous un appariement au cadre interlocutif (les emplois de *moi* avec des éléments subséquents appartenant au cadre délocutif, comme dans *moi c'est dans la boîte / moi c'est la boîte d'oeufs*, sont moins répandus : cf. Apothéloz 1997).

C'est ici qu'intervient, en termes d'approche méthodologique des productions, la sélection de variables généralisables et l'exigence d'une caractérisation des éléments instanciés. En la matière, des entretiens menés à la suite d'analyses de pratiques auprès de professeurs des écoles (Nice, 11-2007 ; 02-2008) montrent que les premières démarches évaluatives que les intervenants mettent en œuvre pour *l'oral* prennent appui sur le nombre d'interventions, combiné éventuellement avec celui des mots et des items distincts. Plus marginalement, les « valeurs » que prennent les verbes fléchis, les syntagmes nominaux, les opérateurs et leur répartition peuvent intervenir, mais cette démarche analytique n'apparaît que chez 2 à 3 pour cent des personnes consultées, ce qui s'explique surtout par le fait des conditions de travail.

Comme il s'agit d'estimer les éventuelles proximités entre cette première analyse et une analyse linguistique plus aboutie, l'un des enjeux de la recherche a ainsi consisté à déterminer l'appropriété des valeurs pressenties, mais aussi, dans le même temps, l'opportunité même de tels corpus instantanés.

3 PREMIER TRAITEMENT DES DONNÉES

3.1 PRÉSENTATION DE LA DÉMARCHE

A la suite d'une première analyse lexicométrique (« atelier lexical » dans Lexica), les regroupements de variables ont été effectués vers des tableaux croisés, avec les variables nominatives dans la première colonne (renvoyant aux enfants), et les variables descriptives dans la première ligne (renvoyant aux proportions de productions), à partir desquelles apparaissent ensuite les valeurs⁷³. Le tableau croisé intègre ainsi un fichier en « données externes » (la première ligne indiquant les variables fermées), dont on demande une analyse factorielle qui correspond plus exactement à une ACP (analyse en composantes principales)⁷⁴. L'ACP prend alors en compte les données descriptives et devient significative, à savoir que, pour ce traitement « 1 » et sur le même instantané de référence, ces données sont les suivantes : les interventions (spontanées et sollicitées), les mots en données brutes et les mots distincts (récurrents ou non), avec les proportions reportées ci-après.

Pour le Groupe 1 :

⁷³ Il suffit d'enregistrer l'ensemble sous la forme d'un tableau croisé en recodant les variables en « fermées multiples », pour reproduire ensuite les données, éventuellement en les transposant sous la forme d'un tableau xls.

⁷⁴La manipulation consiste ici à se rendre sur l'onglet « analyses », puis sur « tableau croisé » : la boîte de dialogue permet ensuite de mettre la variable 1 en colonne et les autres variables « en ligne », en cliquant sur « regrouper dans un même tableau ».

	INTERV	MOTS	MOTS DIST	
APOLLINE		9%	8%	7%
LEO		21%	19%	22%
MARIE SARAI		11%	10%	8%
MATHILDE		11%	10%	13%
MATTEO		22%	23%	22%
NOEMIE		26%	30%	28%

Pour le Groupe 2 :

	INTERV	MOTS	MOTS DIST	
ALAIN		6%	6%	7%
AMBRE		27%	27%	22%
ANTONY		4%	3%	4%
KYLLIAN		15%	14%	15%
LUCAS		13%	15%	15%
MAXIM		20%	25%	21%
OPHELIO		15%	10%	16%

Pour le Groupe 3 :

	INTERV	MOTS	MOTS DIST	
ALAIN		32%	36%	27%
JAED		17%	14%	22%
MATHIS		8%	10%	12%
SARAH		23%	31%	27%
SHON		20%	9%	12%

Les valeurs indiquées correspondent aux proportions individuelles de production chez les enfants. Ainsi Mattéo, dans le premier groupe, accapare-t-il 22 pour cent des interventions parmi l'ensemble de celles relevées chez les enfants, là où les productions de Sarah, par exemple, occupent 31 pour cent des mots produits par le groupe 3. Ces proportions s'inscrivent, comme nous l'avons précédemment indiqué, dans le format d'une analyse multivariée (plusieurs variables sont rassemblées). Or, bien qu'il existe des corrélations entre les trois types de proportions, celles-ci ne sont toutefois pas automatiques (Cf. les exemples d'Ophélio et de Shon, dans les groupes 2 et 3).

Dans tous les cas, l'une des premières questions à soumettre au traitement consiste à estimer dans quelle mesure une représentation, pour ainsi dire liminaire, de ces données, devient productive à partir de ces « variable(s) en ligne » et « variable(s) en colonne ». Ces résultats donnent effectivement des informations succinctes, en ce qu'ils ne font que confirmer des seuils plus ou moins élevés de dispersion du groupe, et déterminent en partie les premiers « écarts-type » entre les productions que les enseignants pressentent assez spontanément. Par ailleurs, ces premières données sont proprement cumulatives (interventions, mots et mots distincts), et ne concernent que les productions verbales en données brutes. En appliquant quoi qu'il en soit les variables en échelles⁷⁵, on peut obtenir un tableau de moyennes caractéristique⁷⁶ : l'analyse multivariée résume alors des tendances statistiques qui organisent des variables descriptives facilement convertibles.

L'ACP a pour mérite d'indiquer le positionnement des variables descriptives les unes par rapport aux autres (représentées ci-après sous forme de vecteurs). Pour le groupe 2 par

⁷⁵ Les variables « 1 » nominatives restent dans une fermée multiple, mais les autres variables (descriptives) passent de « fermées multiples » à « fermées échelles ».

⁷⁶ Ce qui revient à cliquer sur « analyser » dans la présentation du corpus, puis sur « tm », en croisant 1 et les variables descriptives regroupées.

exemple, on remarque que les écarts entre les vecteurs sont très faibles, alors que deux d'entre eux coïncident dans le cas du premier groupe (avec un écart sensible pour le troisième), et que l'ensemble des variables accusent des écarts importants, ce qui signifie que les données brutes ne sont pas *co-orientées*, comme le dégagent les contributions aux facteurs ci-après⁷⁷.

Pour le Groupe 1 :

	Axe 1 (+69.67%)		Axe 2 (+30.31%)	
CONTRIBUTIONS POSITIVES	INTERV	+46,0%	MOTS DIST	+92,0%
	MOTS	+46,0%		
CONTRIBUTIONS NEGATIVES			INTERV	-3,0%
			MOTS	-3,0%

Pour le Groupe 2 :

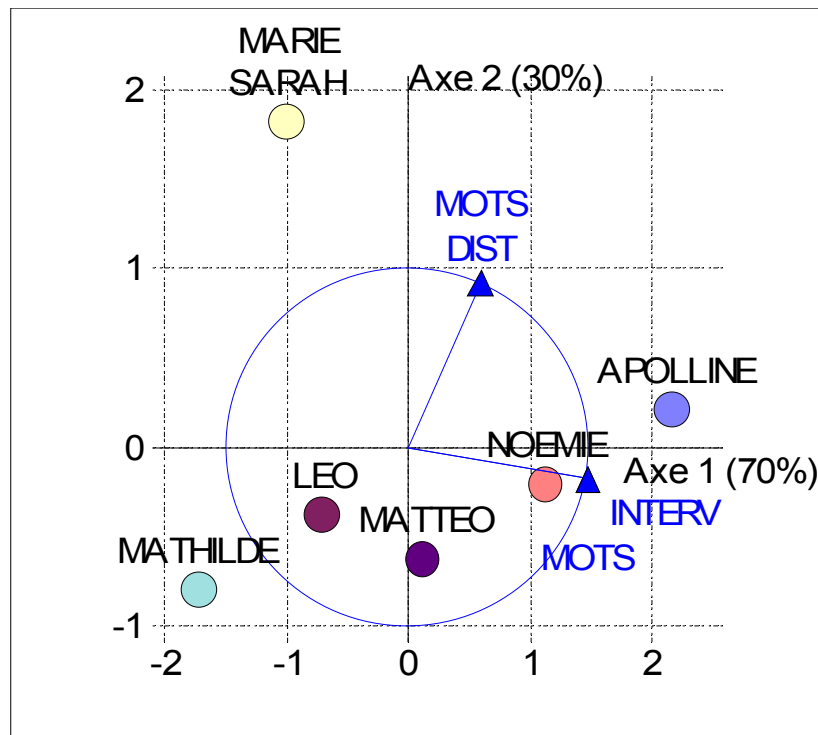
	Axe 1 (+95.40%)		Axe 2 (+4.03%)	
CONTRIBUTIONS POSITIVES	MOTS DIST	+34,0%	MOTS	+67,0%
	INTERV	+33,0%		
CONTRIBUTIONS NEGATIVES			INTERV	-18,0%
			MOTS DIST	-14,0%

Pour le Groupe 3 :

	Axe 1 (+44.63%)		Axe 2 (+33.29%)	
CONTRIBUTIONS POSITIVES	MOTS	+49,0%	MOTS DIST	+76,0%
	MOTS DIST	+11,0%	INTERV	+23,0%
CONTRIBUTIONS NEGATIVES	INTERV	-38,0%		

Plusieurs représentations graphiques existent pour figurer comment se positionnent les variables les unes par rapport aux autres, mais elles n'en sont ni un compte rendu exact, ni encore moins un aboutissement. Si nous prenons le groupe 1 par exemple, avec un taux de variance pour le moins satisfaisant (du fait surtout des faibles quantités de données), une représentation graphique simplifiée place les variables de la manière suivante :

⁷⁷ Dans le tableau de bord constitué, et après avoir recodé les variables descriptives en « fermées échelles », on clique sur « analyses » puis sur « tableau de moyennes / corrélation ». Apparaît une boîte de dialogue. On clique alors sur « pour chaque modalité / valeur de » en mettant « 1 », puis on sélectionne dans les « variables numériques » l'ensemble des variables échelles. Un « groupe 2 » apparaît dans le tableau de bord, en dessous duquel on requiert l'ACP.



L'approche de ce domaine de variabilité interindividuelle impliquant des proportions, elle s'appuie sur des valeurs numériques que les représentations graphiques, à travers l'ACP, projettent sur un plan bidimensionnel dont seuls les segments des vecteurs INTERV, MOTS et MOTS DIST, *a priori*, indiquent la tridimensionnalité (plus ils sont « courts », plus ils se rapprochent du centre). Toutes les valeurs sont ainsi résumées dans le graphique de manière à dégager des composantes principales, lesquelles rassemblent des ensembles de « points » corrélés. Les « contributions aux axes » nous indiquent alors dans quelle mesure telle ou telle variable prend part à la concrétisation des facteurs, laquelle paraît productive ici, vu que plus la somme des taux de variance se rapproche de 100, plus la représentation graphique a des chances d'« expliquer » les liens qui existent entre les variables. Dans le cas reporté ci-dessus, la somme des taux des facteurs 1 et 2 donne précisément 100, ce qui suppose que le graphique explique intégralement les relations qui s'établissent entre les valeurs.

Les tableaux de contributions et les représentations qui leur correspondent confortent quoi qu'il en soit le premier dépouillement pour ce qui relève des variables descriptives. En revanche, elles en facilitent l'abord en ceci qu'elles donnent un peu plus d'indications sur la manière dont se placent les variables nominatives. Or, non seulement il est apparu que les représentations présentent des taux de variance tout à fait satisfaisants sur l'ensemble des groupes, mais qu'elles permettent aussi d'esquisser les moyennes productives en acquisition, en mesurant les écarts quantitatifs et qualitatifs entre les interventions.

Indiquons qu'une autre possibilité, pour ce type d'analyse, est de déterminer les rapprochements éventuels entre les participants, lesquels rapprochements constituent l'un des critères de construction ou de déconstruction du groupe restreint : leur absence marquera éventuellement ce que nous appellerons un domaine de dispersion élevé, alors que plusieurs rapprochements au sein d'un même groupe peuvent marquer une réduction de la dispersion⁷⁸

⁷⁸ Ces derniers éléments sont également envisageables en termes d'insertion, comme en témoigne l'exemple d'« Alain », qui est intégré à la fois dans le deuxième et dans le troisième groupes : dans le premier, il paraît peu inséré, alors que dans le deuxième, son insertion est d'autant plus effective que les écarts sont moindres par

(à condition bien sûr que ces rapprochements ou ces écarts soient validés par la matrice des corrélations et les coordonnées des variables). Sans doute cela apparaîtrait-il davantage avec une application des régressions, que nous ne reprendrons pas ici, mais les calculs de distance sont déjà, de notre point de vue, suffisamment adaptés à cette première démarche.

4 DEUXIEME TRAITEMENT

Afin d'établir dans quelle mesure les enfants opèrent effectivement une structuration discursive de leurs interventions, nous avons mis en oeuvre un deuxième traitement, auquel le premier a été confronté, et qui s'est effectué sur la base d'une analyse linguistique plus aboutie.

Sur les mêmes éléments du corpus de référence, les variables descriptives retenues cette fois-ci ont été les éléments verbaux (éventuellement passivés, négativés ou « amassés » : Gerdes et Kahane 2006), les éléments thématiques et rhématiques (au sens de Mel'čuk 2001), les éléments cadratifs (pour beaucoup déictiques : Charolles *et al.* 2005), et les opérateurs spécifiques (coordonnants et subordonnants en particulier : Torterat 2002, Rebuschi 2002, Desclés 2008), avec les abréviations respectives VERB, THEM, RHEM, CADR et OP. D'autres *témoins* (opérateurs évaluatifs par exemple) auraient pu être pris en compte pour l'étude, pour peu qu'ils eussent été à même de rendre compte d'un processus d'acquisition où leur présence est significative. Pour nous en tenir aux phénomènes de production et de structuration discursives, nous avons caractérisé le corpus non linéaire en nous conformant à l'approche précédente, sans passer par une multi-annotation.

L'autre type de réponse que ce deuxième traitement informatique est appelé à donner revient à définir si une analyse multivariée à l'appui de facteurs linguistiques plus aboutis, confirme en partie, ou non, le premier traitement établi en données brutes, dont les résultats peuvent présenter éventuellement une certaine trivialité, ou du moins se restreindre à la prise en compte de répartitions certes représentatives, mais pour le moins sommaires.

En pratique, la réédition du corpus (en données textuelles balisées) a consisté à étiqueter les éléments linguistiques un par un, en effectuant éventuellement des regroupements locaux, comme c'est le cas par exemple pour les amas verbaux et les syntagmes nominaux autour d'un même noyau prédicatif. Dans le format texte, les verbes et amas verbaux sont indiqués par commodité en italiques (avec éventuellement un trait en bas (*touche 8*) pour spécifier qu'un premier élément forme un tout avec un élément distant), les THEM sont indiqués en majuscules, les RHEM en petites majuscules, les CADR en souligné, et les OP en gras. Voyons ci-dessous trois extraits du corpus caractérisé :

Mathilde : ÇA *se débloque*

Mathilde : ICI

Mathilde : JE *roule* / (JE *roule*)

Mathilde : JE *sais pas*

Mathilde : LA BOITE DES BISCUITS

Mathilde : moi J'ai UNE ROUE **qui marche pas** TRÈS BIEN

Mathilde : moi JE *prends* CETTE BOÎTE

Mathilde : moi TU M'*as donné* UN GRAND // *fais voir* TON BÂTON

Mathilde : OUI JE *veux* UN TROU | LÀ

Mathilde : OUI

rapport aux autres, et qu'un domaine de connivence s'établit.

Mathilde : UN / DEUX

Mathilde : voilà

Lucas : *c'est* CASSÉ

Lucas : *c'est* DRÔLE

Lucas : *il y a* ~~eu~~ // UNE BOÎTE ?

Lucas : JE *prends* CELLE-LÀ

Lucas : JE *prends* LES BRINDILLES | À LA MER **et** LÀ | LES BRINDILLES | À LA MONTAGNE

Lucas : ON *dirait* UNE ÉPÉE // ~~Taya~~

Lucas : OUI *c'est* VRAI **ÇA** *marche*

Lucas : OUI

Lucas : *regarde* **comment** ELLE *est* ma voiture

Lucas : VOITURES

Lucas : ON *joue* À LA COURTE PAILLE ?

Shon : *ata*

Shon : *ata*

Shon : ~~heu~~

Shon : OUI

Shon : VOITURE

Shon : VOITURE

Shon : *roule*

La prise en compte par « regroupements » des éléments d'après Lexica ne permet pas un étiquetage automatique complètement satisfaisant, étant donné qu'il procède par assemblages pour une part fortuits et ne distingue pas certains (pseudo)homographes. De ce fait, nous avons opéré une réédition manuelle sur l'intégralité du corpus. Sans préconstruire en rien la quantification des données et comme indiqué *supra*, le relevé effectué nous a conduit à recouper des informations paradigmatiques, syntagmatiques et topologiques, ainsi que certaines informations contextuelles.

Plusieurs choix ont été effectués pour ce relevé, où l'on remarquera ci-après l'absence d'opérateurs spécifiques (comme les coordonnants et les subordonnants) dans le troisième groupe, et des écarts quelquefois importants, quoique prévisibles, entre les proportions des variables. Par exemple, les verbes et amas verbaux (où nous avons intégré les forclusifs de négation) s'accompagnent d'un *ata* de Shon qui, même si l'on peut estimer qu'il s'agit d'un cadratif grammaticalisé, renvoie à ce que nous avons identifié en contexte, momentanément du moins, comme un *attends* actionnel (Cf. Balthasar *et al.* 2003). D'autre part, il n'a pas été tenu compte, dans le relevé proportionnel, ni des opérateurs suspensifs du type *heu* (qui marquent des opérations diverses et nécessitent une prise en compte des contours prosodiques), ni des interjections plus ou moins onomatopéiques du type *ah*, *aye* ou *iliou* (lesquelles peuvent être plus ou moins thématiques, rhématiques ou cadratives : Wilkins 1992, Cuenca et Hilferty 1999).

Plusieurs occurrences du pronom *il* avec l'unipersonnel *falloir* apparaissent également dans les interventions, mais dans la mesure où, en marge de l'autocorrection qu'il suppose, ce pronom est non thématique sur le plan de la signification discursive, il a été intégré à l'amas

verbal dont il fait partie (cette sélection est bien évidemment discutable, mais les proportions qu'elle occupe dans les productions verbales sont minimales). Nous n'entrerons pas dans le détail des difficultés plus « fines » de cette caractérisation, qui fera l'objet d'autres contributions (Torturat 2010 ; *soum.*), et nous insisterons ici sur le traitement des données, lequel, expérimenté sur les trois groupes (lesquels présentent des variances assez satisfaisantes dans l'analyse), nous informe des proportions suivantes.

Pour le Groupe 1 :

	VBS	THEM	RHEM	CADR	OP
APOLLINE	9,00%	12,50%	5,00%	11,00%	4,00%
LEO	19,00%	16,00%	20,00%	15,00%	13,00%
MARIE SARAI	7,00%	7,00%	11,00%	19,00%	0,00%
MATHILDE	10,00%	13,00%	11,00%	9,00%	4,00%
MATTEO	25,00%	21,00%	24,00%	23,00%	8,00%
NOEMIE	30,00%	30,00%	29,00%	23,00%	71,00%

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%

Groupe 3 :

	VBS	THEM	RHEM	CADR	OP
ALAIN	33,00%	44,00%	6,00%	62,00%	0,00%
JAED	9,00%	14,00%	25,00%	13,00%	0,00%
MATHIS	15,00%	14,00%	19,00%	0,00%	0,00%
SARAH	28,00%	14,00%	44,00%	25,00%	0,00%
SHON	15,00%	14,00%	6,00%	0,00%	0,00%

Ces données font ainsi apparaître qu'Alain, une fois revenu dans le troisième groupe (le sien), accapare 33 pour cent des verbes et amas verbaux du groupe, 44 pour cent des éléments thématiques, mais seulement 6 pour cent des éléments rhématiques. De même dans le groupe 2, Lucas n'emploie que 6 pour cent des cadratifs prédiqués dans le cadre des co-verbalisations, mais ses productions oscillent entre 12 et 15 pour cent des emplois recensés par ailleurs.

Confrontées au même type d'analyse que lors des précédents traitements, ces données démontrent que diverses combinaisons de variables sont possibles (la variable OP étant non représentative pour le Groupe 3, elle est considérée comme « nulle »). Or, les traitements intermédiaires que nous avons pratiqués sur ce corpus discursif ont montré que les analyses accomplies, avec là aussi des taux de variance satisfaisants, impliquent des contributions négatives aux facteurs très diverses selon les groupes, comme en témoignent les bilans ci-dessous.

Pour le groupe 1 :

	Axe 1 (+43.86%)		Axe 2 (+30.62%)	
CONTRIBUTIONS POSITIVES	RHEM	+33,0%	OP	+58,0%
	VBS	+28,0%	RHEM	+13,0%
CONTRIBUTIONS NEGATIVES	CADR	-31,0%	THEM	-9,0%
	THEM	-6,0%	VBS	-6,0%

Pour le groupe 2 :

	Axe 1 (+53.88%)		Axe 2 (+25.65%)	
CONTRIBUTIONS POSITIVES	OP	+33,0%	RHEM	+52,0%
	CADR	+24,0%	VBS	+17,0%
CONTRIBUTIONS NEGATIVES	RHEM	-6,0%	CADR	-14,0%

Pour le groupe 3 :

	Axe 1 (+60.40%)		Axe 2 (+24.86%)	
CONTRIBUTIONS POSITIVES	CADR	+35,0%	VBS	+54,0%
	THEM	+32,0%	CADR	+1,0%
CONTRIBUTIONS NEGATIVES			RHEM	-42,0%
			THEM	-1,0%

Dans le même temps, les matrices de corrélations s'avèrent particulièrement informatives, ce qui est le cas par exemple pour le groupe 1, où les contributions négatives indiquent une certaine dispersion des variables relevées :

	C1	C2	C3	C4	C5
C1 : VBS	1,000				
C2 : THEM	0,143	1,000			
C3 : RHEM	0,589	-0,327	1,000		
C4 : CADR	-0,617	0,327	-0,471	1,000	
C5 : OP	-0,151	-0,151	0,520	0,384	1,000

Variance expliquée par les composantes :

	f1	f2	f3	f4	f5
Valeur propre	2,193	1,531	1,053	0,170	0,052
% expliqué	43,866%	30,629%	21,063%	3,404%	1,038%
% cumulé	43,866%	74,495%	95,558%	98,962%	100,000%

Les représentations graphiques et les matrices de corrélations confirment le fait que de telles combinaisons de variables ne permettent pas de distinguer ce qui, dans les productions des jeunes enfants, porte spécifiquement sur les manières dont ils construisent leur discours et répondent de manière organisée aux sollicitations de l'intervenant, ou aux interventions des autres participants. Une analyse en composantes principales permettant de tester plusieurs combinaisons de variables, de manière à considérer les indices de variance les plus satisfaisants, il nous a donc paru opportun de combiner d'un côté les éléments VBS, RHEM et OP, puis de l'autre les éléments THEM, CADR et OP, avec des représentations que nous

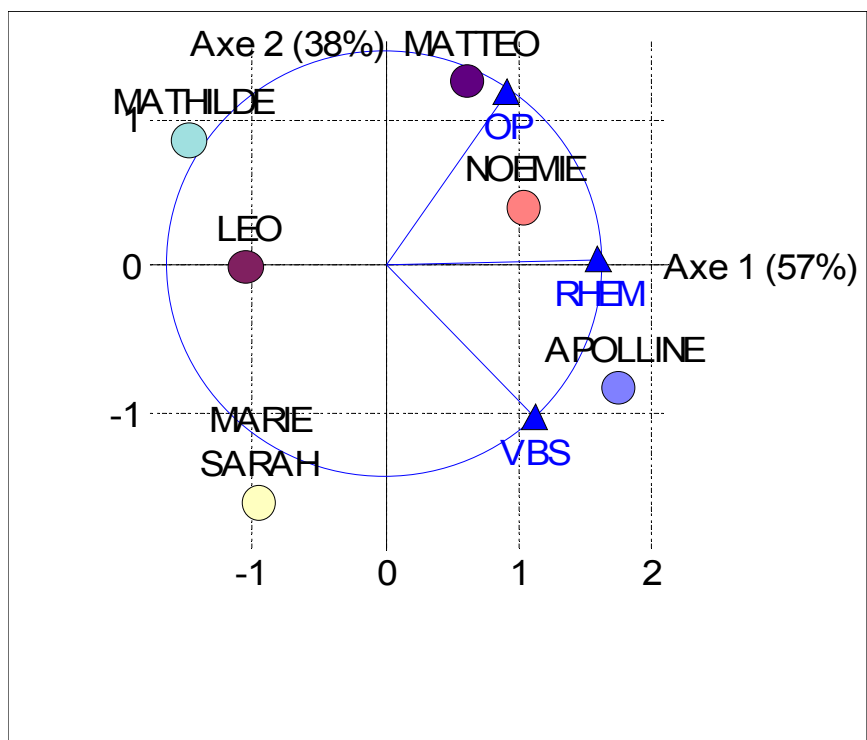
reporterons ci-après uniquement pour le premier groupe, où les productions sont pour le moins diverses.

Pour justifier ce choix en quelques lignes, disons que l'emploi des verbes et amas verbaux et celui des éléments rhématiques ont pour point commun, parmi d'autres, de ne pas porter principalement sur les manières dont les enfants encadrent, spécifient ou présentent ce à propos de quoi les informations sont données. Généralement non focaux, les VBS et RHEM constituent la plupart du temps des apports en lien avec des éléments précédemment prédiqués, ce qui n'est pas le cas de ce qui relève de la thématique (THEM) et de l'encadrement (CADR) discursifs, lesquels coïncident le plus souvent avec des éléments qui apparaissent dans la zone préverbale. Bien entendu, la plus grande mesure s'impose dans ce domaine, et il ne s'agit pas de relever ces récurrences pour en faire autre chose que des tendances, mais ces dernières permettent toutefois de donner plus de visibilité aux démarches discursives des enfants, lesquelles démarches, à l'oral, se combinent avec de multiples autres indicateurs, au premier rang desquels interviennent les phénomènes relatifs aux données prosodiques et à la corporéité⁷⁹.

Indiquons que l'intégration de la variable des opérateurs (OP) dans les deux combinaisons de variables descriptives s'explique à partir de deux postulats simples : d'une part, les opérateurs sont des marqueurs de structuration syntagmatique, phrastique et discursive et, en tant que tels, sont proprement des polyopérateurs, en ceci qu'ils interviennent dans bien des cas sur plusieurs dimensions simultanément (Culioli 1990). De ce fait, même en les discriminant à l'appui d'indications co(n)textuelles précises, il est très difficile d'en garantir complètement le classement. D'autre part, ces opérateurs segmentaux posent des difficultés propres à leur intervention dans le cadre de corpus oraux, dans ce sens où ils s'accompagnent à l'oral de marqueurs suprasegmentaux (intonatifs : Berrendonner (2004) ou gestuels : Bouvet (2001), par exemple). Or, ces derniers se combinent avec eux et en réduisent la portée singulière.

Dans le cas du Groupe 1, les regroupements de variables apportent des informations intéressantes, dans la mesure où elles dégagent des tendances qui laissent une place à une première interprétation des données. Ci-après pour le premier regroupement (VBS, RHEM et OP) :

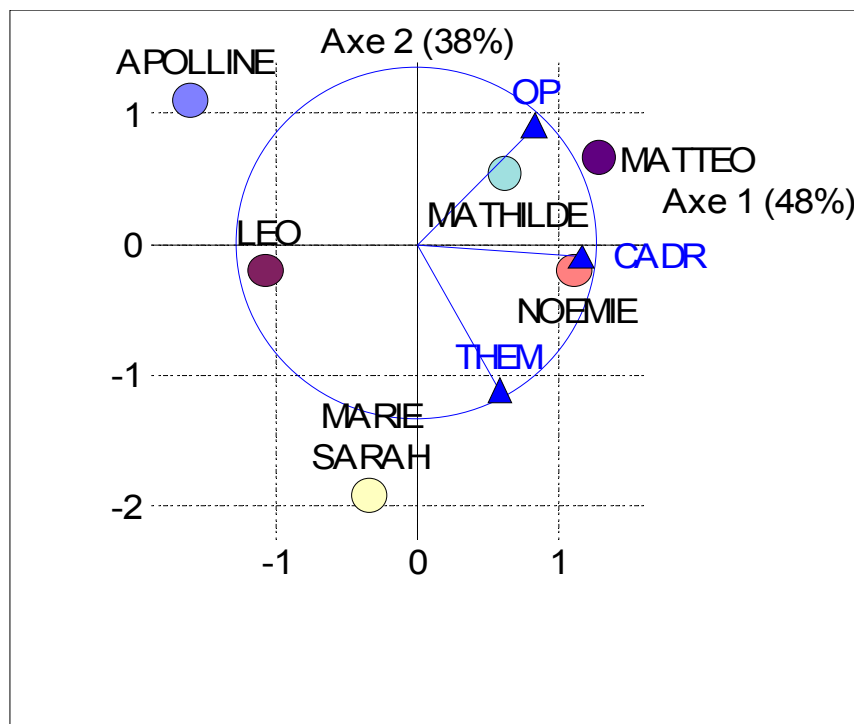
⁷⁹ *Corporéité* entendu comme « ensemble des traits concrets du corps comme être social », d'après les termes de Berthelot (1983). Ces traits sont principalement posturo-mimo-gestuels.



Contribution aux facteurs :

	Axe 1 (+57.14%)	Axe 2 (+38.31%)
CONTRIBUTIONS POSITIVES	RHEM +54,0%	OP +56,0%
	VBS +26,0%	
CONTRIBUTIONS NEGATIVES		VBS -43,0%

Pour le deuxième regroupement (THEM, CADR et OP) :



Contribution aux facteurs :

	Axe 1 (+47.87%)	Axe 2 (+38.33%)
CONTRIBUTIONS POSITIVES	CADR +57,0%	THEM +58,0%
	OP +27,0%	
CONTRIBUTIONS NEGATIVES		OP -40,0%

Ces représentations ainsi que les autres, par combinaisons diverses de variables, nuancent en partie la trivialité du premier traitement en données brutes, mais confirment dans le même temps ses insuffisances. Elles révèlent par ailleurs que les faiblesses, ou au contraire l'abondance des productions dans les domaines de la catégorie verbale et des éléments rhématiques masquent en partie celles qui relèvent proprement de la structuration discursive.

Dans le premier regroupement, on note que le facteur 1 rassemble les contributions des éléments rhématiques et dans une moindre mesure des éléments verbaux, et que le facteur 2 privilégie les opérateurs, au détriment notamment des éléments verbaux. Dans le deuxième regroupement, le facteur 1 rend principalement compte des cadratifs et des opérateurs, alors que le facteur 2 privilégie les éléments thématiques.

Les positionnements des variables nominatives ne peuvent se comprendre qu'à partir de ces éléments contributifs, lesquels sont donc diversement explicatifs sur le plan statistique. Ils emportent néanmoins avec eux ce que nous appellerons volontiers des *suggestions*, comme dans le cas de Marie-Sarah, laquelle, en dépit de l'emploi significatif qu'elle fait des cadratifs, mais aussi du fait qu'elle verbalise autant que Mathilde et plus encore qu'Apolline, voit ses productions en partie à l'écart de celles du noyau du groupe, à la fois en termes de production et de structuration discursives.

Cela n'est vraisemblablement pas le cas, en l'occurrence, d'Apolline et de Mathilde, qui en termes de proportions dans le groupe, se positionnent différemment selon les variables expliquées. Or, nous retrouvons ici l'absence notamment d'une « métrique simple » des composantes principales, lesquelles ne permettent pas, *a priori*, de formuler des déductions robustes à partir des représentations graphiques basées sur des instantanés.

5 QUELQUES CONCLUSIONS

Les conclusions que nous tirerons des applications présentées *supra* sont donc surtout d'ordre méthodologique. La première est que, dans le cadre des analyses multivariées de ce type de corpus discursif sous forme d'ACP, les positionnements des variables nominatives, dans les représentations graphiques, renvoient plus à des profils de production qu'à des profils d'acquisition, même si l'ACP permet de rendre plus concrets les rapprochements et les écarts entre les variables descriptives, auxquels elle apporte des garanties numériques. La deuxième est que l'ensemble de ces éléments confortent en partie le point de vue suivant lequel il convient de distinguer ce qui concerne les productions en données brutes par rapport à ce qui concerne spécifiquement la structuration discursive. En la matière, le choix qui consiste à regrouper d'un côté les verbes et amas verbaux avec les éléments rhématiques, et de l'autre les éléments thématiques avec les cadratifs, à partir du moment où le traitement appliqué au corpus étiqueté fait l'objet d'une analyse multivariée, suppose de multiplier les tests pour déterminer quelles peuvent être les différentes contributions des opérateurs.

En marge du coût opérationnel que représentent ces types de traitements, la principale objection que l'on pourrait formuler ici est que l'analyse multivariée appliquée sur des corpus instantanés demeure tout à fait incomplète si elle n'est pas confortée par une version longitudinale. En effet, les déductions qu'il est possible d'envisager à partir d'instantanés, non seulement se révèlent intermédiaires, mais aussi nécessitent des recoupements constants dans les informations statistiques et leurs corrélats informatiques, avec tout ce que les représentations graphiques correspondantes comportent d'apriorisme.

Les corpus longitudinaux, en relatant les productions verbales de groupes restreints d'enfants collectées sur plusieurs mois ou plusieurs années, en plus du fait qu'ils apportent davantage d'indications dans le domaine de la composition du lexique (Bassano, Eme et Champaud 2005), contribuent également à mesurer les acquisitions des enfants à travers la « grammaire » qu'ils mettent en oeuvre et l'organisation discursive qui s'établit dans leurs interventions. D'autre part, de par leur volume et leur diversification, ils font l'objet d'un suivi régulier et permettent de valoriser des tendances et des constantes qui ne peuvent être que pressenties à partir de corpus instantanés.

Ces derniers nous semblent toutefois productifs à quelques égards. En effet, les instantanés permettent de pratiquer une analyse exploratoire du corpus, et d'effectuer une première approche de données textuelles pour une part converties en variables numériques. En outre, ils conduisent à mesurer en partie la représentativité du corpus et à en aborder diversement la matérialité. Ce type d'apport s'avère par conséquent plus opportun au moment du tri parmi les démarches analytiques possibles, qu'à celui du traitement informatique des données. A ce titre, ils représentent quand même un gain de temps non négligeable, d'autant qu'ils contribuent à dégager ce qui, dans les démarches envisagées, concerne plus particulièrement les difficultés de mise en oeuvre.

6 RÉFÉRENCES

- Apothéloz D. (1997). « Les Dislocations à gauche et à droite dans la construction des schématisations », dans : A. Berrendonner et D. Miéville (éds.), *Logique, Discours et Pensée. Mélanges offerts à Jean-Blaise Grize*, Bern, Lang : 183-217.
- Balthasar L., Bruxelles S., Mondada L, Traverso V. (2003). « Attends ça fait travailler le cerveau : usages et tendances à la grammaticalisation de *attends* en français parlé en interaction », dans : *Linguistique de Corpus, 36e colloque de la Societas Linguistica Europea*, Lyon, 4-7.
- Bassano D., Eme E., Champaud C. (2005). « A naturalistic study of early lexical development : General processes and inter-individual variations in French children ». *First Language*, 25 (1), 67-101.
- Benzitoun C. (2004). « L'Annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? », dans : *Actes du Colloque TALN 2004*, URL : <http://aune.lpl.univ-aix.fr/jep-taln04/proceed/actes/recital2004/Benzitoun.rec04.pdf> [consulté le 25 avril 2007].
- Berrendonner A. (2004). « Grammaire de l'écrit vs grammaire de l'oral : le jeu des composantes micro- et macro-syntaxiques », dans : A. Rabatel (éd.), *Interactions orales en contexte didactique*, Lyon, PUL : 249-264.
- Berthelot J.M. (1983). « Corps et Société. Problèmes méthodologiques posés par une approche sociologique du corps ». *Cahiers internationaux de sociologie*, LXXXIV : 119-131.
- Bouvet D. (2001). *La Dimension corporelle de la parole*. Louvain, Peeters.
- Charolles M., Le Draoulec A., Pery-Woodley M.P., Sarda L. (2005). « Temporal and spatial dimensions of discourse organisation ». *Journal of French Language Studies*, 15-2 : 203-218.
- Cibois P. (2007). *Les Méthodes d'analyse d'enquête*. Paris, PUF.
- Cuenca M.J., Hilferty J.J. (1999). *Introducción a la lingüística cognitiva*. Barcelona, Editorial Ariel.
- Culioli A. (1990). *Pour une linguistique de l'énonciation*, 1. Paris, Ophrys.
- Desclés J.P. (2008). « Opérations de prédication et de détermination ». *Lidil*, 37 : 61-98.
- Dister A., Simon A.C. (2008). « La Transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé », *Arena Romanistica* 1/1 : 54-79.
- Gerdes K., Kahane S. (2006). « L'Amas verbal au cœur d'une modélisation topologique de l'ordre des mots ». *Linguisticae Investigationes*, 29-1 : 75-89.
- Grobet A. (1997). « La Ponctuation prosodique dans les dimensions périodique et informationnelle du discours », *Cahiers de Linguistique française* 19 : 83-123.
- Mac Whinney B. (2000). *The CHILDES project : Tools for analysing talk* (troisième édition). Mahwah NJ, Lawrence Erlbaum.
- Martin O. (1997). « Aux origines des idées factorielles. Des théories aux méthodes statistiques ». *Histoire & Mesure*, 12 (3-4).
- Mel'čuk I. (2001). *Communicative Organization in Natural Language*. Amsterdam/Philadelphia : Benjamins.
- Pincemin B. (2007). « Introduction » au numéro 6 de la revue *Corpus* : 5-15 [article consulté le 6 juillet 2009] : <http://corpus.revues.org/index812.html> .
- Rastier F. (2009). « Quantité et Qualité en sémantique de corpus ». *Comm. aux 6èmes Journées de Linguistique de corpus*, Lorient, Université de Bretagne-Sud.
- Rebuschi G. (2002). « Coordination et subordination : vers la co-jonction généralisée ». *Bulletin de la Société de Linguistique de Paris*, 97-1 : 37-94.
- Sansonetti L. (2003). « Approche lexicométrique de corpus d'interactions verbales entre un adulte et un enfant en cours d'acquisition du langage. Résultat d'expérience » : dans G. Williams (éd.), *Actes des Troisièmes Journées de la Linguistique de Corpus*, Lorient, Université de Bretagne-Sud [consulté de 30 mai 2009] : http://web.univ-ubs.fr/corpus/jlc3/1_4_sansonetti.pdf .
- Saussure L. de, Sthioul B. (2002). « Interprétations cumulative et distributive du connecteur *et* : temps, argumentation, séquencement ». *Cahiers de linguistique française*, 24 : 293-314.
- Shriberg E., Stolcke A., Hakkani-Tur D., Tur G. (2000). « Prosody-Based Automatic Segmentation of Speech into Sentences and Topics », *Speech Communication* 32-1 : 127-154.

- Sinclair J. (2005). « Meaning in the Framework of Corpus Linguistics », dans : W. Teubert (éd.), *Lexicographica*, Tübingen, Niemeyer : 20-32.
- Tortérat F. (2002). *Approche des invariants de quelques joncteurs en français : pour une complémentarité des termes de coordination et de jonction*. Paris-Sorbonne, thèse de doctorat.
- Tortérat F. (2010). « Le Recours aux corpus discursifs : difficultés et possibilités pratiques », dans : M. St. Rădulescu (éd.), *La Méthodologie pour un apprentissage de la recherche*, Chisinau, Université de Ion Creangă.
- Tortérat F. (soum.). « Analyse de quelques variabilités interindividuelles dans le domaine de la structuration discursive : à partir d'un corpus caractérisé », dans : *Actes du Colloque « Linguistic Approaches to Text Structuring »*, Paris, ENS.
- Wilkins D.P. (1992). « Interjections as deictics ». *Journal of Pragmatics* 18 / 2-3 : 119-158.