Apports de la statistique textuelle pour le repérage des reprises et reformulations dans les corpus d'interaction verbale entre un adulte et un enfant

Luiggi Sansonetti

EA 2290 SYLED, EA 170 CALIPSO – Université de la Sorbonne Nouvelle, Paris 3
19 rue des Bernardins – 75005 Paris – France
luiggi.sansonetti@noos.fr

Abstract

Besides the functions of quantification of linguistic units and comparison of use and evolution of the lexicon by author, by genre or by period, textual statistics allows to track down easily the phenomena of repetitions and reformulations in language acquisition processes in adult-child dialogue corpora.

Résumé

En plus de ses fonctions de quantification d'unités linguistiques et de comparaison d'emploi et d'évolution du lexique par auteur, par genre ou par période, la lexicométrie permet de repérer facilement les phénomènes de reprises et de reformulations dans des corpus de dialogues entre un adulte et un enfant en cours d'acquisition du langage.

Mots-clés : statistique textuelle, lexicométrie, analyse de corpus, acquisition du langage, interaction verbale adulte/enfant

1. Introduction

La linguistique de l'acquisition recherche les processus d'élaboration du fonctionnement cognitivo-langagier, en observant tout particulièrement la syntaxe chez l'enfant, à partir d'interactions verbales avec un adulte, en situation de parole spontanée. Cette approche s'appuie sur les avancées du courant interactionniste (Ochs et Schieffelin, 1995) et des travaux sur le français parlé (Blanche-Benveniste, 1997).

Dans un corpus longitudinal de dialogues entre un adulte et un enfant, nous pouvons observer des phénomènes de reprises et de reformulations de la part des deux locuteurs. Lorsque l'enfant reprend à l'adulte une construction et l'emploie dans une formulation inadéquate, et que l'adulte reformule cette construction de manière adéquate, l'adulte répond aux tâtonnements de l'enfant et lui permet alors la vérification de son hypothèse. C'est ce que Wyatt appelle « feed-back correctif » (Wyatt, 1969), désigné aujourd'hui par « interaction ajustée ou adaptée ».

Ces interactions verbales sont enregistrées puis transcrites selon des conventions de transcription propres aux objectifs de recherche. Les deux corpus utilisés dans le cadre de cet exposé ont suivi les conventions de transcription établies par Lentin et ses collaborateurs (Lentin *et al.*, 1998). Le premier corpus longitudinal, nommé « Julien », est comprend trois dialogues entre une même adulte et un même enfant, entre 5 ans 10 mois et 6 ans 4 mois. Le deuxième

corpus longitudinal, nommé « Mathilde », comprend aussi trois dialogues entre le même adulte et un autre enfant, entre 4 ans 9 mois et 4 ans 11 mois (Tissier, 2001). Pour ces interactions, l'adulte a utilisé un livre illustré *Crictor* (Ungerer, 1958) pour analyser les caractéristiques des interactions adulte-enfant dans la co-construction d'une narration.

2. Analyses « traditionnelles » et analyses informatisées

Dans le cadre de l'analyse des catégories syntaxiques d'énoncés, en particulier pour déceler l'approbation de constructions complexes par l'enfant, à partir de la transcription orthographique des dialogues enregistrés, nous recherchons dans les énoncés de l'adulte et de l'enfant des mots *(quand, puisque)*, des groupes de mots (*il faut que*, *parce que, pour que*), des constructions syntaxiques *(verbe + verbe infinitif)* ainsi que des tentatives de constructions syntaxiques chez l'enfant. Pour chaque dialogue, une grille d'analyse est remplie. Une synthèse présente un récapitulatif des éléments et des constructions trouvés, les références des énoncés contenant ces éléments, avec leurs fréquences pour chaque dialogue. L'évolution des fréquences de chaque catégorie syntaxique dans le corpus longitudinal est ensuite représentée par des courbes.

Le Projet CHILDES[1] qui présente un ensemble de programmes (CLAN) pour pratiquer diverses analyses sur des corpus de dialogues (MacWhinney, 2000), permet la recherche de formes et/ou de morphèmes, ainsi que des répétitions de formes. Cependant, il ne s'agit que d'une recherche énoncé par énoncé. Avec le logiciel de lexicométrie et de statistique textuelle Lexico3[2], il est possible de travailler énoncé par énoncé mais aussi sur l'ensemble des dialogues en même temps.

3. Repérage de « pourquoi – parce que »

Pour analyser les corpus avec Lexico3, nous avons suivi des « phases de préparation » (Sansonetti, 2003), pendant lesquelles nous avons notamment ajouté des « clés », permettant ainsi de partitionner le corpus longitudinal en dialogues ou en locuteurs (Lebart et Salem, 1994). De plus, nous avons regroupé les énoncés 50 par 50, d'abord ceux de l'adulte, puis ceux de l'enfant, et ainsi de suite. De cette façon, nous obtenons une cartographie du corpus de Julien répartie par dialogues avec sur les lignes impaires les énoncés de l'adulte, et sur les lignes paires ceux de l'enfant.

3.1. *Le corpus de Julien*

À l'aide de la fonction « Segments Répétés » de Lexico3, nous dressons la liste des suites de formes dont la fréquence est supérieure à 2. Nous recherchons, par exemple, l'emploi de la locution « parce que » et sa répartition dans le corpus longitudinal de « Julien ».

Nous constatons que le premier « parce que » est énoncé par l'enfant dans l'énoncé 20 du premier dialogue (ju=120)[3]. Cette tentative de l'enfant n'est pas reprise par l'adulte, mais par l'enfant lui-même, à l'énoncé ju=121 dans une construction syntaxique complète :

<ad=121> et pourquoi tu aimes bien ce livre ?
<ju=120> ***parce que***
<ad=122> tu aimes bien les serpents ?

[1] http://childes.psy.cmu.edu
[2] http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/
[3] Selon les conventions de transcription, seuls les énoncés ne contenant que « mm » ne sont pas numérotés. C'est pourquoi, dans nos corpus informatisés, il peut y avoir un décalage dans la numérotation des énoncés.

*<ju=121> euh non mais c'est ce que je / c'est **parce que** j'avais envie*

Quand nous recherchons la motivation de ce « parce que » chez l'enfant, nous nous apercevons qu'il est toujours en réponse à un « pourquoi » de l'adulte :

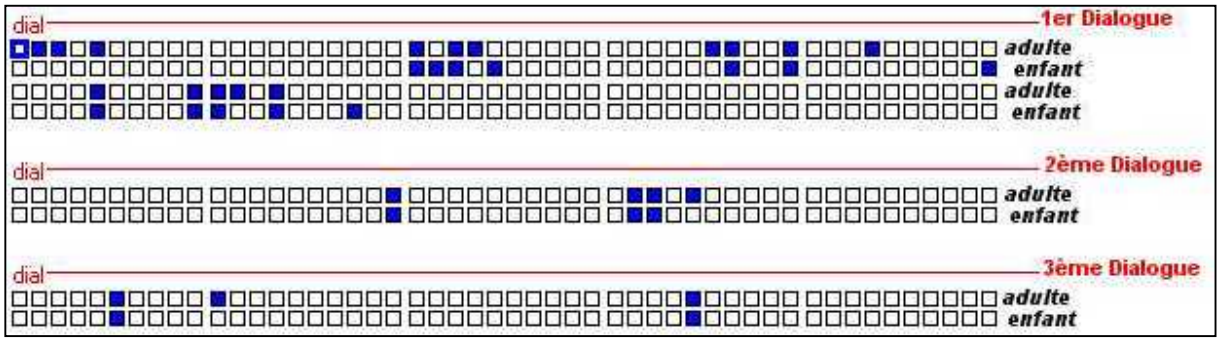

Figure n° 1. Cartographie des « pourquoi » et de « parce que » dans les corpus de Julien.

Chaque petit carré représentant un énoncé, il est facile de revenir au texte pour en vérifier le contenu. Nous avons réuni dans un même groupe de formes les « pourquoi » et « parce que ». Les carrés colorés indiquent la présence dans l'énoncé d'un élément du groupe de formes. Lorsque nous recherchons seulement les « pourquoi », nous remarquons, que l'enfant ne pose jamais ce type de question.

3.2. *Le corpus de Mathilde*

Lorsque nous procédons à la même recherche dans le corpus longitudinal de « Mathilde », nous remarquons qu'il n'y a aucune question de type « pourquoi ».

Le premier « parce que » de l'enfant, seulement énoncé dans le deuxième dialogue, est une tentative abandonnée qui n'est pas reprise immédiatement par l'adulte :

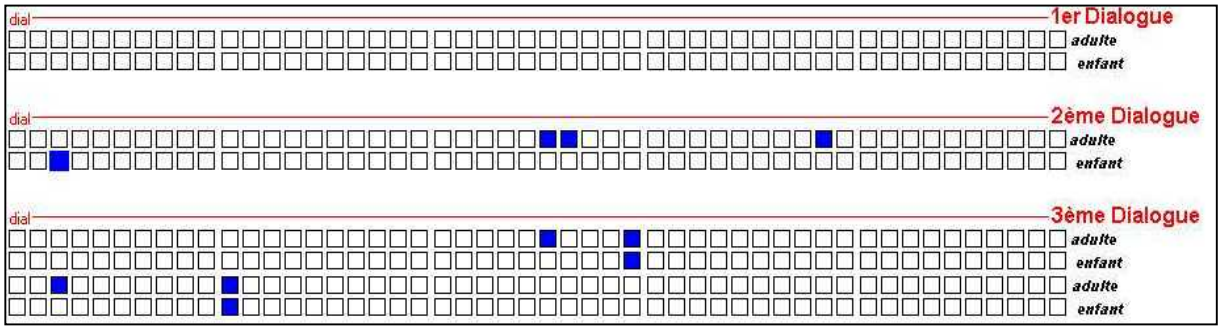

Figure n° 2. Cartographie des « parce que » dans les corpus de Mathilde

*<ma=203> elle va au zoo **parce que** le serpent trictor alors elle l'appela Crictor alors*

Dans ce troisième corpus, l'adulte énonce plusieurs constructions avec « parce que » que l'enfant ne reprend pas immédiatement. Il n'y a que trois productions de « parce que » par Mathilde dont les deux dernières, dans le troisième dialogue, sont des reprises immédiates des énoncés de l'adulte :

*<ad=329> **parce que** le mot néant commence par la lettre N*
*<ma=330> d'accord **parce que** / **parce que** heu dans / dans quoi*
[...]
*<ad=360> [...] il a un parc qui porte son nom **parce que** il a / il a / il a arrêté le cambrioleur*

<ma=360> d'accord ***parce que*** *il a arrêté le cambrioleur*

4. Repérage de catégories et de structures syntaxiques

Après avoir étiqueté les corpus « Julien » et « Mathilde » par Cordial4, nous reprenons les recherches de reprises et reformulations, non plus de formes, mais de constructions syntaxiques. Nous recherchons par exemple les constructions de type « Préposition + Verbe Infinitif » (que nous désignerons par « Prep+VInf »).

4.1. *Julien-Forme-Cat*

Dans le corpus de « Julien-Forme-Cat » (les formes sont suivies des catégories grammaticales), lorsque nous regroupons toutes les constructions « Prep+VInf » qui apparaissent au moins 2 fois dans le corpus, nous obtenons la liste suivante :

à_PREP compter_VINF	*6*
à_PREP faire_VINF	*3*
à_PREP manger_VINF	*2*
pour_PREP voir_VINF	*3*

À partir de ce groupe de formes, considéré comme une seule entité, nous recherchons les reprises de ces constructions pour en vérifier les contextes dans le corpus. Nous observons alors une production d'une construction « à+VInf » de cette structure par l'enfant après un verbe conjugué de façon non grammaticale.

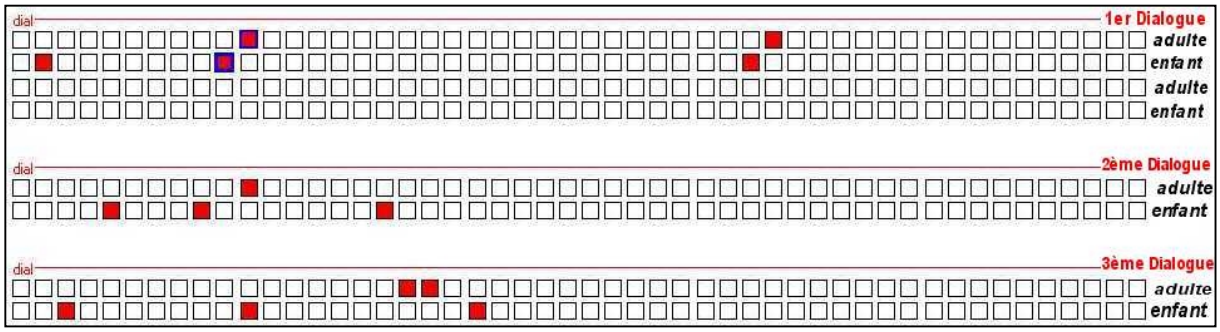

Figure n° 3. Cartographie des « Prep+VInf » dans le corpus étiqueté de Julien.

L'adulte va immédiatement reformuler l'énoncé de l'enfant avec une morphologie verbale grammaticale en utilisant le passé composé :

<ju=109> il_PPER3S apprena_NCI ***à_PREP compter_VINF***
<ad=111> donc_COO il_PREP a_VINDP3S il_PPER3S a_VINDP3S appris_VPARMS ***à_PREP compter_VINF*** *à_PREP l'_DETDFS école_NCFS aussi_ADV*

4.2. Mathilde-Forme-Cat

Dans le corpus de « Mathilde-Forme-Cat », nous remarquons que le nombre d'occurrences de la structure « Prep+VInf » est le double de celui du corpus de Julien (14 énoncés au total pour le corpus de Julien, 27 pour celui de Mathilde). Nous remarquons également avec la cartographie du corpus que les énoncés sont principalement regroupés en paires.

[4] http://synapse.com

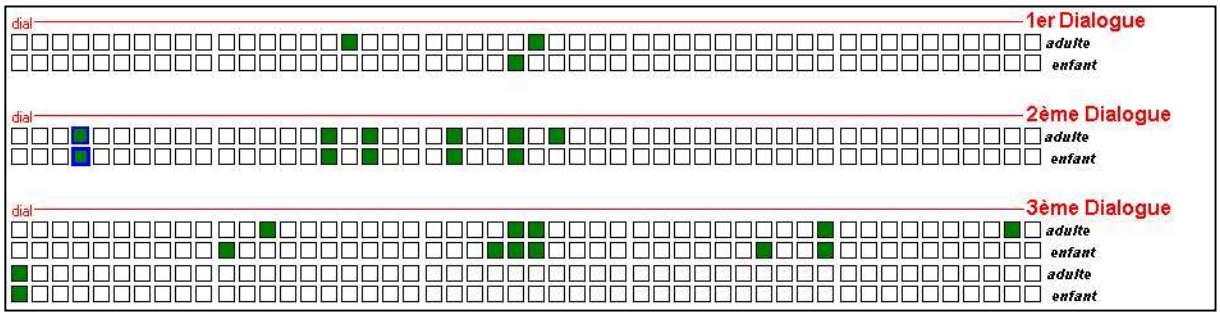

Figure n° 4. Cartographie des « Prep+VInf » dans le corpus étiqueté de Mathilde.

4.3. *Problèmes d'étiquetage automatique*

Lors de l'étiquetage de nos corpus, nous savions que les modèles de langage utilisés par Cordial sont basés sur du texte, à l'origine, écrit, et qu'il n'est pas adapté au langage parlé. Bien des études ont traité des problèmes de l'étiquetage automatique de données orales[5].

Dans un des exemples précédents, nous avons le cas de formes inconnues propres à la création de l'enfant sur le modèle des verbes du premier groupe comme « apprena ». Dans les paramétrages de l'étiqueteur, nous avons donné l'instruction de ne pas corriger la faute. L'étiqueteur l'a bien repérée mais n'a rien fait, puisque, dans notre étude qui peut être aussi lexicale ou morphologique, nous voulions conserver la forme énoncée par l'enfant. Ainsi dans le temps, il serait possible de voir si l'enfant, à force ou non de recevoir des constructions au passé simple et/ou des corrections adaptées de la part de l'adulte, passe de « apprena » ou « metta » à « apprit » et « mit ».

Une autre erreur d'étiquetage plus délicate qui s'est produite est le fait d'avoir plusieurs étiquettes ou aucune pour les mêmes formes ou mêmes locutions. Dans le corpus « Julien-Forme-Cat », nous avons *pour_PREP* et *pour qu'_SUB*. Alors que dans le corpus « Mathilde-Forme-Cat », nous avons *pour_PREP, pour qu'/que_SUB* et *pour_NCMS*. Même si pour les conjonctions de subordination, le *pour* n'est pas étiqueté isolément, ceux qui sont notés NCMS (nom commun masculin singulier) sont dus à un article précédant la préposition :

<ad=337> […] six comme les ***pour*** *les six pattes d'un insecte […]*

5. Le corpus *Julien-Mathilde*

Puisqu'il s'agit du même adulte dialoguant avec Julien puis Mathilde, nous réunissons les deux corpus en un seul, ce qui permet comparer les enfants entre eux, ainsi que l'adulte à elle-même. Nous confirmons le fait qu'il y a presque le double d'occurrences de la construction *« Prep+VInf »* chez « Mathilde » par rapport à « Julien ».

En faisant un graphique de ventilation pour une partition donnée, il est possible d'observer l'évolution de l'emploi de la forme dans les corpus, ici dans le temps. Nous obtenons le graphique suivant :

[5] Nous pouvons citer les travaux de Valli et Véronis (1999), Paroubek et Rajman (2000) et Véronis (2000).

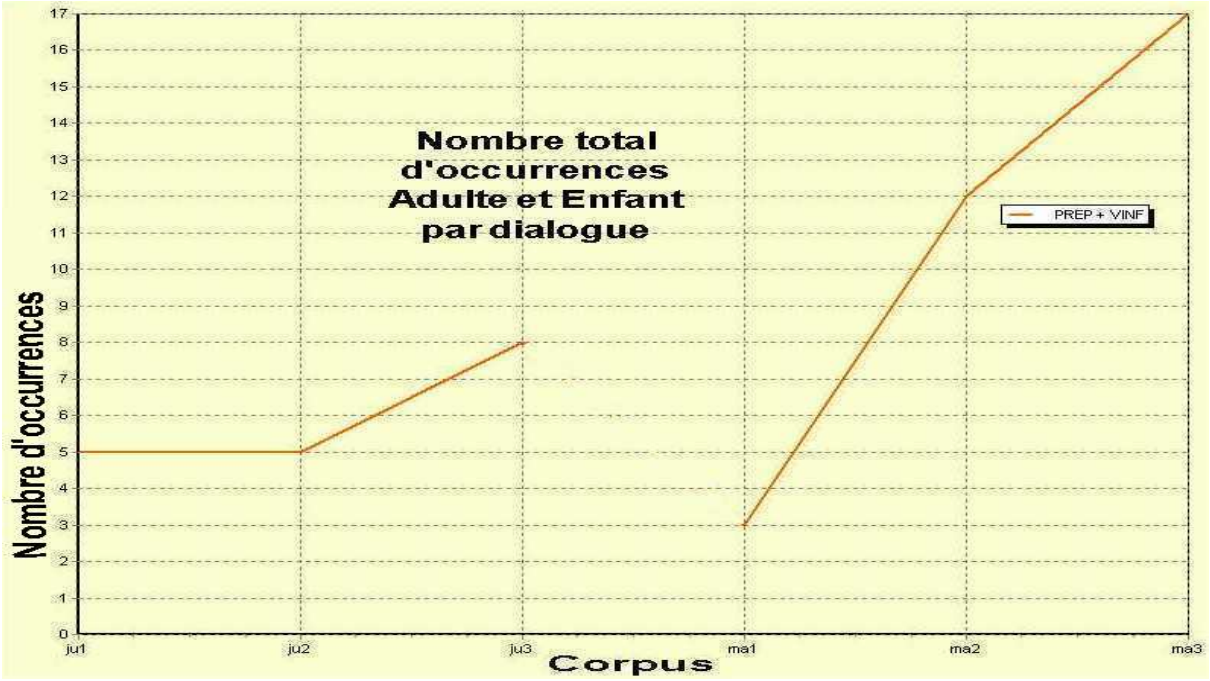

Figure n° 5. Comparaison de l'évolution de la construction « Prep+VInf » dans les corpus étiquetés de Julien et Mathilde réunis.

Nous constatons que dans les dialogues avec Mathilde, il y a de plus en plus d'occurrences de cette construction. Lorsque nous sélectionnons la partition « locuteurs », nous pouvons alors observer dans la figure suivante, dans un premier temps, que le nombre d'occurrences de cette construction chez l'adulte et l'enfant correspondant (ad1 avec ju1 et ad2 avec ma2) suit la même évolution. Nous pouvons observer, dans un deuxième temps, que Mathilde a eu plus d'offres de « Prep+VInf » que Julien de la part de l'adulte.

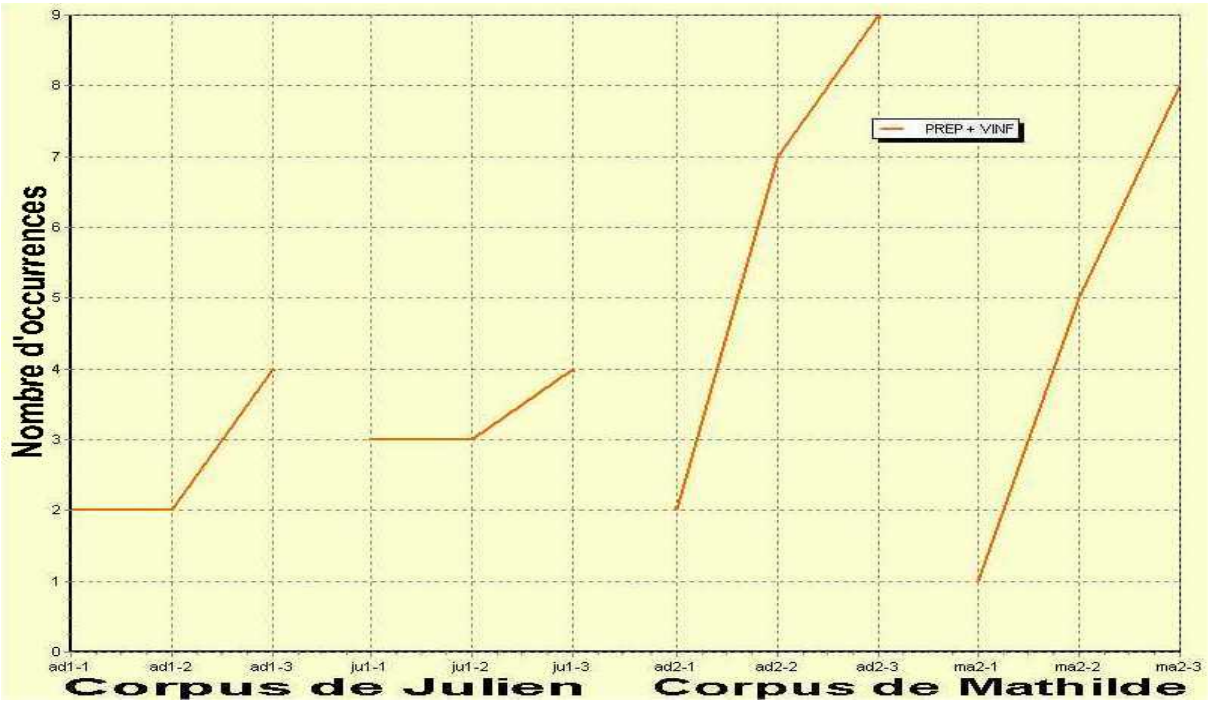

Figure n° 6. Comparaison de l'offre de la construction « Prep+VInf » de l'adulte et de sa reprise par chaque enfant dans les corpus étiquetés de Julien et Mathilde réunis.

6. Conclusion

À l'aide de logiciels de statistique textuelle et d'étiquetage automatique, il est possible d'étudier des données orales et d'observer l'évolution des emplois du vocabulaire et de la syntaxe. Il est également possible avec le logiciel Lexico3, après des phases de préparation du corpus, de visualiser rapidement les reprises et les reformulations qui sont des phénomènes pertinents dans le processus d'acquisition de la langue en situation de dialogue.

Il est montré par notre expérience que plus l'adulte propose de mots, des constructions différentes et des réutilisations dans des contextes différents, plus l'enfant essaie de reproduire puis de produire ces éléments de façon spontanée. C'est le cas de Mathilde pour les constructions du type « Prep+VInf ». Bien sûr, il faudra revenir sur les résultats des étiquetages automatiques de données orales, en nous laissant la possibilité de corriger les étiquettes avant deuxième traitement.

Références

Blanche-Benveniste C. (1997). *Approches de la langue parlée en français*. Ophrys, Collection l'essentiel français.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Lentin L. *et al.* (1984). *Recherches sur l'Acquisition du Langage*, t. (1). Presses de la Sorbonne Nouvelle.

Lentin L. *et al.* (1988). *Recherches sur l'Acquisition du Langage*, t. (2). Presses de la Sorbonne Nouvelle.

Lentin L. (1997). *Apprendre à penser, parler, lire, écrire*. ESF.

MacWhinney B. (2000). *The CHILDES project. Tools for analyzing talk* (2 vol.). Mahwah, Lawrence Erlbaum Associates.

Ochs E. et Schieffelin B. (1995). Language socialization and its consequences for language development. In Fletcher P. et MacWhinney B. (Eds), *The Handbook of Child Language*. Blackwell Publishers : 73-94.

Paroubek P. et Rajman M. (2000). Étiquetage morpho-syntaxique. In Pierrel J.-M. (Ed.), *Ingénierie des langues*. Hermès Sciences Publications : 131-150.

Sansonetti L. (2003). Approche lexicométrique de corpus d'interactions verbales entre un adulte et un enfant en cours d'acquisition du langage. Résultats d'expérience. Actes du Colloque « 3es Journées de la Linguistique de Corpus », Lorient, 11, 12 et 13 Septembre 2003. Presses Universitaires de Rennes (à paraître).

Tissier C. (2001). *Rôle de l'adulte dans l'interaction langagière adulte-enfant (entre 4 ans 9 mois et 6 ans 4 mois) en situation de narrations dans deux corpus longitudinaux*. Mémoire de Maîtrise, ILPGA, Paris 3 (non publié).

Ungerer T. (1958). *Crictor*. L'Ecole des Loisirs pour l'édition en langue française, 1980, réédition 2000. Collection lutin poche.

Valli A. et Véronis J. (1999). Étiquetage grammatical des corpus de parole : problèmes et perspectives. *Revue Française de Linguistique Appliquée*, vol. (IV/2) : 113-133.

Véronis J. (2000). Annotation automatique de corpus : panorama et état de la technique. In Pierrel J.-M. (Ed.), *Ingénierie des langues*. Hermès Sciences Publications : 151-171.

Wyatt G. (1969). *La relation mère – enfant et l'acquisition du langage*. Mardaga.