

# Journée d'étude

## « Corpus oraux : problèmes méthodologiques de recueil et d'analyse de données »

*Nancy - vendredi 27 mars 2009*

**ATILF Bâtiment CNRS**

*Campus Lettres et Sciences Humaines - 44 avenue de la libération*

**Coordination : Tiphane Bertin et Magali Husianycia  
(Laboratoire ATILF/CNRS Nancy-Université)**

**Comité scientifique :**

Virginie André, Christophe Benzitoun, Tiphane Bertin, Emmanuelle Canut, Jeanne-Marie Debaisieux, Magali Husianycia et Evelyne Jacquey.

**Comité d'organisation :**

Delphine Beuseroy, Iveta Chovanova, William del-Mancino, Laurent Gobert, Aurore Koehl, Sabrina Martin, Sandrine Pescarini, Sandra Tomc, Jean-Marc Voirin, Crépin Yaouédéou.

Contact : [magali.husianycia@atilf.fr](mailto:magali.husianycia@atilf.fr)

Site : [www.atilf.fr](http://www.atilf.fr)

## PRÉSENTATION

Depuis une dizaine d'années, de nombreux colloques et autres rencontres ont abordé la problématique des corpus oraux : recueil de données orales et transcriptions, bonnes pratiques, analyses « outillées » de grand corpus, archivage, catalogage, codage, etc. L'évolution des technologies a permis d'accéder à de nouvelles pratiques dans le travail de constitution et d'analyse de corpus. Il existe aujourd'hui de nombreux logiciels : des logiciels d'aide à la transcription, essentiellement orthographique (Transcriber) tantôt couplée à de l'analyse multimodale (Transana), tantôt couplée à une analyse de la voix (Praat), des logiciels d'assistance à l'analyse quantitative et/ou qualitative (Unitex). D'autres logiciels sont utilisés, cependant, ces derniers ne sont pas spécialisés dans le traitement des données orales (comme par exemple Lexico 3, Unitext ou encore les logiciels d'étiquetage comme Tree Tagger).

L'objectif de cette journée d'étude est de coordonner une confrontation des différentes approches sur les données orales et d'aborder les problèmes méthodologiques de recueil et d'analyse de ces données orales :

- Déterminer le type de corpus à recueillir en fonction des objectifs de recherche,
- Trouver un terrain d'enquête,
- Définir les modes d'enregistrement,
- Déterminer les modalités de transcription,

Tout en sachant que les problèmes varient selon les objectifs des chercheurs.

Il s'agit de mettre l'accent sur la méthodologie de l'analyse, sur les entrées de travail possibles à partir d'un corpus et sur la détermination des unités d'analyse en fonction des objectifs de recherche :

- Quelles analyses des données orales ?
- Comment analyser les données ?
- Quels sont les logiciels nécessaires pour l'analyse ?

Cette journée d'étude doit être d'une part l'occasion de faire l'inventaire des pratiques et outils existants et d'autre part de mener une réflexion sur la méthodologie d'analyse.

## PROGRAMME

8h30-9h00

Accueil

9h00-9h15

**Mot de bienvenue** : Jean-Marie Pierrel, Directeur de l'ATILF  
**Présentation de la journée** : Tiphany Bertin et Magali Husianyca

*Présidente de séance : Virginie André*

9h15-10h15

**Analyser un corpus de langue parlée en interaction : questions méthodologiques**

Véronique Traverso (Université de Lyon)

10h15-10h30

Pause

10h30-11h30

**Des données représentatives...de quoi en acquisition du langage ?**

**Constitution de données à observer et objectifs d'analyse**

Martine Vertalier (Université de la Sorbonne nouvelle - Paris 3)

Emmanuelle Canut (ATILF/CNRS - Nancy-Université)

11h30-12h30

**Les corpus, la diversité, les variétés, la variation**

Françoise Gadet (Université de Paris Ouest Nanterre la Défense)

12h30-14h00

Déjeuner

*Présidente de séance : Jeanne-Marie Debaisieux*

14h00-15h00

**La transcription et ses entours**

Paul Cappeau (Université de Poitiers - FoReLL)

15h00-16h00

**De l'intérêt des corpus diversifiés pour les descriptions en syntaxe**

Mireille Bilger (Université de Perpignan-via-Domitia)

16h00-16h15

Pause

16h15-17h15

**Réflexions en vue d'élaborer une grammaire du français parlé sur corpus. La question des données.**

Christophe Benzitoun (ATILF/CNRS - Nancy-Université)

17h15-17h30

Conclusion

José Deulofeu (Université de Provence - Aix Marseille 1 - DELIC)

# RÉSUMÉS DES COMMUNICATIONS

## Analyser un corpus de langue parlée en interaction : questions méthodologiques

**Véronique Traverso (Université de Lyon)**

Dans cette présentation, nous nous attacherons à deux questions essentiellement : les liens entre l'établissement du corpus et les questions de recherche d'une part et d'autre part les deux méthodologies essentielles d'abord des corpus en analyse d'interaction (analyse longitudinale d'un cas ; établissement de sous-corpus autour d'un phénomène).

La première de ces deux questions conduira à mettre en perspective l'ensemble des étapes conduisant de l'observation des situations sociales à l'analyse d'un corpus. La seconde illustrera deux façons d'entrer dans l'analyse, à partir de deux corpus très différents dans leur dimension et leur nature.

## Des données représentatives... de quoi en acquisition du langage ? Constitution de données à observer et objectifs d'analyse

**Martine Vertalier (Université de la Sorbonne nouvelle - Paris 3)  
Emmanuelle Canut (ATILF/CNRS - Nancy-Université)**

Nous aborderons la question de la constitution et de l'analyse des données dans le domaine de l'acquisition du langage en tentant de faire ressortir les différences entre une approche psycholinguistique et une approche linguistique de l'étude du langage de l'enfant de moins de six ans. Pour cela, nous partirons d'articles d'un numéro récent de la revue *Journal of Child Language* (revue considérée comme la plus représentative de l'avancée des recherches en acquisition du langage dans le champ de la psycholinguistique). Dans un premier temps, à partir du recensement des termes employés par les auteurs pour évoquer les locuteurs, les données et les méthodes de recueil, de sélection, de « codage » et d'analyse des données, nous poserons les questions suivantes :

- Quelles sont les données recueillies : le langage ou le comportement de qui ? Quel langage ? Quelle durée ? Recueilli par qui ? Comment ? Dans quelles situations ?
- Quelles sont les données sélectionnées pour l'analyse ? Comment sont-elles transcrites, « codées » ? Comment sont-elles analysées ?

Dans un deuxième temps, nous comparerons les modalités de ces travaux en psycholinguistique aux objectifs poursuivis par l'approche qualitative des recherches en linguistique de l'acquisition.

## Les corpus, la diversité, les variétés, la variation

**Françoise Gadet (Université de Paris Ouest Nanterre la Défense)**

Cet exposé se place dans une perspective qui voudrait documenter des faits de variation du français, sur le plan syntaxique. Il présentera donc une réflexion, en amont de la constitution de corpus, sur ce qui se trouve à la source de la diversification linguistique : est-ce que ce sont bien les catégories socio-démographiques dont relèvent les locuteurs, comme le suppose bien souvent la pratique actuelle de collecte de corpus, qui s<sup>1</sup>inscrit ainsi implicitement dans une sorte de sociolinguistique simplifiée, mettant en avant la co-variation ?

On défendra une position plus complexe, afin de montrer que tout choix méthodologique a des conséquences théoriques (avec des effets pervers si on ne les prend pas en compte à la source), et que les phénomènes (en particulier pour la syntaxe) ne sont pas tous à considérer de la même manière.

## La transcription et ses entours

**Paul Cappeau (Université de Poitiers - FoReLL)**

Le fort développement que connaissent les projets liés aux corpus oraux rend utile de revenir sur les enjeux liés à la transcription. L'intervention, qui tire parti d'une expérience déjà ancienne dans la transcription, sera construite autour des 4 questions suivantes :

- 1) Quelles sont les pratiques actuelles en ce domaine et quels sont les points sur lesquelles elles se différencient ?
- 2) Quels sont les types d'erreurs que commettent les transcrip-teurs débutants ? Comment intervenir sur leur formation ?
- 3) Quelles hypothèses ces écarts dans la transcription permettent-ils d'envisager ?
- 4) Quelles seraient les conséquences de l'exploitation de corpus oraux non corrigés ?

### Quelques références

- Baude, Olivier (éd.). 2006. *Corpus oraux - Guide des bonnes pratiques*. Paris. CNRS Editions.
- Bilger, Mireille (éd.). 2008. *Données orales - Les enjeux de la transcription*. Perpignan. PUP.
- Blanche-Benveniste, Claire & Jeanjean, Colette. 1986. *Le français parlé. Édition et transcription*. Paris. Didier-Erudition
- Blanche-Benveniste, Claire. 1997. *Approches de la langue parlée en français*. Paris. Ophrys.
- Bond, Zinny S. 2005. "Slips of the Ear", dans Pisoni, David B. & Remez, Robert E. (eds). 2005. *The Handbook of Speech Perception*. Malden. Blackwell Publishing. 290-310.
- Cappeau, Paul. 2008. "Perception et reconstruction" dans Bilger, Mireille (éd.). *Données orales : les enjeux de la transcription*. Perpignan. PUP. 235-247.

## De l'intérêt des corpus diversifiés pour les descriptions en syntaxe.

**Mireille Bilger (Université de Perpignan-via-Domitia)**

Si le débat sur l'utilité de travailler à partir de données attestées semble clos, il n'en demeure pas moins vrai qu'il convient de rester vigilant face à ce consensus qui peut cacher certaines divergences, entre autres, sur la fonction accordée au corpus (réservoir d'exemples vs matériau source pour la description) ou sur la « pertinence » du corpus étudié. De fait, la portée - ou encore la validité - de la description basée sur ces données risque d'être fort différente selon la façon dont a été conçu le corpus et les objectifs pour lesquels ce dernier l'a été.

En ce qui concerne le domaine (morpho)-syntaxique, nombreux sont les travaux qui illustrent l'intérêt de s'appuyer sur des corpus variés et échantillonnés afin d'en parfaire la description, d'en renouveler l'analyse ou d'en proposer une nouvelle présentation. Lors de notre communication nous rappellerons certains de ces résultats (études sur le pronom *lequel*, la conjonction *et*, etc.), ce qui nous permettra de revenir sur des oppositions souvent présentées comme fondamentales, telles que celles que l'on a pu poser entre oral et écrit, lexique et grammaire, ou encore entre « système » et « usages » du système.

### Quelques références

- BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S. & FINEGAN, E., 1999, *Longman grammar of spoken and written English*. London: Pearson.
- BIBER, D., 2006, *University Language. A corpus-based study of spoken and written registers*. Amsterdam. John Benjamins Publishing Company.
- BILGER, M. et CAPPEAU, P. « Les Données des corpus ou comment dépasser certaines représentations » in Abecassis & alii (actes du colloque *Les voix du français : usages et représentations* », AFLS, 3-5 septembre 2008, Oxford)
- CAPPEAU, P. et GADET, F., 2007, « Maître-mot et pierre philosophale: l'exploitation sociolinguistique des grands corpus ». *RFLA*, vol XII-1: 99-110.
- CORI, M., DAVID, S., 2008, "Les corpus fondent-ils une nouvelle linguistique ?". *Langages*, n° 171, 111-129.
- HABERT, B., 2000, "Des corpus représentatifs : de quoi, pour quoi, comment ?" dans Bilger, M. (éd). « Linguistique sur corpus - Etudes et réflexions », *Cahiers*, n° 31, P.U. de Perpignan, 11-58.
- HALLIDAY M.A.K., 1985, *Spoken and written Language*. Oxford. Oxford University. Press. HALLIDAY, M.A.K., 1991, "Corpus studies and probabilistik grammar », in Aijmer, K. & Altenberg, B. (eds), *English Corpus Linguistics*, Londres-Ney-York, Longman, 30-43.

## Réflexions en vue d'élaborer une grammaire du français parlé sur corpus. La question des données.

**Christophe Benzitoun (ATILF/CNRS - Nancy-Université)**

Sous ce titre extrêmement ambitieux se cache en fait un travail préliminaire très limité. Cela fait quelques années que périodiquement l'idée de poser les bases d'une grammaire du français parlé sur corpus se manifeste, ce qui m'a amené, avec l'aide appuyée de mes collègues scientifiques, à réfléchir sur la faisabilité d'un tel projet et à faire un certain nombre de propositions allant dans ce sens. Pourtant, à ce jour, aucun article scientifique n'a été finalisé par manque - c'est mon hypothèse - d'une réflexion soutenue et ciblée sur les prémices d'une telle entreprise. Et malheureusement, par le passé, les ouvrages portant sur la description du français parlé sur corpus ont été relativement rares, ce qui évidemment ne facilite pas notre tâche. Nous en dressons ci-dessous une liste non exhaustive et peu détaillée.

Martinon (1927), dans son ouvrage intitulé *Comment on parle en français*, a mené une description du français parlé à visée normative, et ce malgré les contraintes techniques de l'époque. Les travaux de Bauche (1928) et Frei (1929) s'intéressaient plutôt à une langue non normative et pas spécialement au français parlé, ce qui ne nous intéressera pas directement ici. Cependant, ces deux productions pourraient nous indiquer quelques pistes intéressantes. Puis, dans les années cinquante, ce fut la constitution du *Français Fondamental*, qui a essentiellement donné lieu à des listes de vocabulaire. Dans les années 60, l'*Enquête Sociolinguistique à Orléans* a fourni essentiellement de nombreux enregistrements et transcriptions. De nos jours, les travaux de Blanche-Benveniste (et collègues) et de Gadet proposent des descriptions ciblées et systématiques, mais sans avoir pour ambition une visée exhaustive.

On le voit bien à l'énumération de ces travaux, une entreprise de grammaire du français parlé sur corpus reste encore un objectif difficilement atteignable et pour lequel données, outils et méthodes restent largement à inventer<sup>1</sup>. Pourtant, on aurait pu s'attendre à ce que la situation s'améliore avec l'omniprésence de l'informatique censée faciliter le travail de collecte et d'exploitation des données orales. Cela constitue d'ailleurs un paradoxe : aucune description grammaticale d'ensemble du français parlé n'a été lancée malgré l'arrivée des ordinateurs, des enregistreurs numériques, etc. et malgré l'importance d'un tel enjeu pour l'ensemble de la société.

Notre présentation se fera sous la forme de questions et de propositions. Malgré le caractère extrêmement ambitieux de la tâche, il nous semble indispensable de lancer au plus vite la réflexion sur des bases aussi précises que possibles. Dans cette perspective, nous essaierons de faire un état des lieux des contraintes à prendre en compte, des besoins et des biais à éviter. Sans avoir l'ambition d'égaliser la *Longman Grammar of Spoken and Written English*, nous espérons juste savoir si le travail que nous menons a une chance d'aboutir dans des délais raisonnables ou si, compte tenu du contexte, de nombreuses années seront nécessaires pour commencer à lancer une telle initiative. Cette journée d'étude me paraissait être le lieu idéal pour lancer une telle réflexion et profiter des lumières des personnes présentes.

---

<sup>1</sup> Il existe bien évidemment de nombreux travaux à visée méthodologique portant par exemple sur la constitution et l'exploitation des données orales, mais sans avoir pour finalité celle que nous décrivons ici.

**Impression**  
ATILF / CNRS – Nancy-Université  
44, avenue de la Libération  
BP 30687  
54063 Nancy Cedex