

## **AUTOUR DU CORPUS DE RÉFÉRENCE DU FRANÇAIS PARLÉ**

**Recherches sur le Français Parlé, vol. 18, 2004, Publications de l'Université de Provence, Aix-en-Provence, ISBN 2-85399-571-2 — 24 €**

*Note de lecture de Jean-Yves Antoine*

*Mél : Jean-Yves.Antoine@univ-tours.fr*

---

Comme bien d'autres langues, le français parlé a longtemps été ignoré des linguistes qui lui préféraient l'étude de l'écrit. Avant l'invention du magnétophone, la difficulté de recueil de corpus oraux fiables a pu justifier ce manque d'intérêt. Depuis, il est clair que cette désaffection de la communauté scientifique ne trouve sa justification que dans une vision faussement dévalorisée de l'oral. Pour nombre de syntacticiens, la langue parlée ne serait ainsi qu'un langage fautif par rapport à une norme écrite qui représenterait une compétence langagière seule digne d'étude. Pour le cogniticien, il est pourtant évident que ces prétendues fautes de l'oral (répétitions, reprises, amorces, lapsus, etc.) sont précisément le reflet des activités cognitives qui président à une production langagière en direct. D'un point de vue strictement linguistique, ces « inattendus » sont pour la même raison des révélateurs très précieux du fonctionnement et de la logique des systèmes langagiers. On saisit dès lors l'importance de l'étude sur la langue parlée qui est sortie, au cours des années 80-90, du purgatoire dans lequel elle était confinée.

En ce qui concerne le français parlé, Claire Blanche-Benveniste et ses collègues du Groupe Aixois de Recherches Syntaxiques (GARS) ont beaucoup fait pour cette reconnaissance. Le groupe DELIC, qui lui a fait suite sous la direction de Jean Véronis, poursuit cette entreprise en s'appuyant plus fortement sur l'utilisation de l'outil informatique. Le groupe DELIC vient ainsi de terminer la constitution d'un corpus oral informatisé d'envergure — le *Corpus de Référence du Français Parlé ou CRFP* — qui constitue le pendant du corpus — essentiellement papier — *CorpAix* qu'avait développé le GARS sur plusieurs décennies. Par sa taille (440 000 mots correspondant à 36 heures de parole), le CRFP n'a pas d'équivalent en France et ne peut se comparer qu'avec les corpus francophones Valibel (Belgique) et d'Ottawa-Hull. Le dix-huitième volume de la collection « Recherches sur le français parlé » fait un premier bilan de l'utilisation de ce corpus d'importance qui sera diffusé au sein de la communauté scientifique<sup>1</sup>.

Le CRFP est tout d'abord présenté de manière détaillée dans un texte co-signé par l'ensemble du groupe DELIC. Outre un rappel des conventions de transcription qui reprennent à quelques exceptions celles définies par le GARS, ce texte décrit les principes d'échantillonnage (répartition géographique des locuteurs, situations de parole, âge, sexe et niveau scolaire des locuteurs, etc.) qui ont présidé à la constitution du corpus. Ce travail d'échantillonnage témoigne d'un réel effort pour

---

<sup>1</sup> Pour plus d'informations sur les modes de diffusion du corpus, on consultera le site Internet du groupe DELIC : <http://www.up.univ-mrs.fr/delic/crpf>.

rendre le corpus représentatif. Il est cependant clair que le CRFP ne peut encore atteindre le statut de corpus de référence unique et autosuffisant en matière de français parlé (ceci est-il d'ailleurs possible?). Tout d'abord, le nombre d'enregistrements qu'il comprend (134) limite encore sa représentativité en termes distributionnels. Mais surtout, il est regrettable qu'une situation de parole soit sous représentée dans le CRFP : il s'agit du dialogue interactif, qui ne peut être comparé aux situations de parole publique (émissions radiophoniques, entretiens) réunies dans le corpus. Cette absence me semble d'autant plus regrettable qu'elle restreint l'usage du corpus pour des études pragmatiques. D'autre part les situations de dialogue, avec leurs interruptions et leur co-construction dynamique du sens présentent des indices très intéressants, à mon sens, pour l'étude des phénomènes de production langagière.

Quoi qu'il en soit, on ne saurait faire grief au DELIC de ces limitations. Lorsqu'on connaît la difficulté et la lourdeur de la transcription de corpus oraux, on comprend aisément que des choix répondant aux priorités du groupe aient présidé à la constitution de ce corpus. On ne peut donc que souhaiter que d'autres centres de recherches poursuivent cet effort pour combler les lacunes des ressources orales francophones. Ce fut précisément l'objectif de l'Action Spécifique ASILA<sup>2</sup> et celui que je poursuis, pour le dialogue oral, avec le projet PAROLE PUBLIQUE<sup>3</sup>.

Après cette première présentation, l'ouvrage dresse un panorama relativement exhaustif des études qui ont déjà été menées sur cette ressource. Un article un peu à part (Marie-Noëlle Roubaud) s'attache à décrire les problèmes rencontrés par les transpositeurs en présence d'amorces. On peut le recommander aux chercheurs et étudiants qui n'ont jamais « affronté » de près la langue parlée. Cet article leur donnera en effet une idée de la difficulté que constitue la transcription, aussi objective que possible, des corpus oraux.

À l'opposé, les autres articles recueillis dans ce volume s'attardent sur des études linguistiques relativement variées sur le corpus transcrit. Je ne vais pas ici discuter de chacun des thèmes abordés : le lecteur intéressé trouvera, à la fin de cette revue critique, le sommaire détaillé de l'ouvrage pour s'en faire une idée plus précise. Disons simplement que ce panorama des premières recherches menées sur le CRFP (mais également sur le corpus *CorpAix*) montre tout l'intérêt scientifique de cette ressource :

- Je noterai tout d'abord que les études qui nous sont présentées relèvent bien entendu de la syntaxe mais également de la pragmatique (article de Catherine Chanet sur les marqueurs discursifs à l'oral) et de la prosodie (contribution d'Estelle Campione sur l'interaction entre pauses silencieuses et pauses remplies). Sur ce dernier point, il n'est peut-être pas inutile de rappeler que la transcription du CRFP est alignée avec le son<sup>4</sup>.

---

<sup>2</sup> Cf. <http://www.loria.fr/projets/asila/>

<sup>3</sup> Cf. [http://www.sir.blois.univ-tours.fr/~antoine/parole\\_publicue/](http://www.sir.blois.univ-tours.fr/~antoine/parole_publicue/)

<sup>4</sup> De mon point de vue, cet alignement ne doit concerner que le mode de diffusion avec fichiers sons au format WAV et transcriptions au format XML avec DTD Transcriber. Il est dommage que le texte de présentation du corpus ne soit pas plus précis sur ce type d'information technique.

- Par ailleurs, ce CRFP se prête à des approches méthodologiques variées. Il peut ainsi servir à étayer une théorie par la « simple » observation d'exemples attestés dans le corpus. Cette approche qualitative est parfaitement illustrée par l'article de Sandrine Candéo sur les détachements avec pronoms disjoints (structures *lui+SN* et *SN+lui*). Par le choix convainquant d'exemples attestés, l'auteur montre en quoi la découverte en corpus de certaines observations peut conduire à des théories novatrices parfois contre intuitives.

À l'opposé, la majorité des recherches présentées dans cet ouvrage reposent sur des études quantitatives parfois très fouillées. Si ces articles donnent des éclairages intéressants sur des phénomènes précis, je me permettrai d'émettre une réserve sur les études statistiques qui sont présentées. Le corpus CRFP a été constitué pour atteindre une certaine représentativité distributionnelle. Ainsi, dans son article sur les adverbes en *-ment*, Mireille Bilger reprend le plan d'échantillonnage du corpus pour établir des différences d'usage en situations de parole privée, professionnelle ou publique. Néanmoins, la plupart des auteurs s'affranchissent, rapidement de mon point de vue, de toute précaution méthodologique dans l'exposé de leurs données statistiques. La discussion repose en effet essentiellement sur des comparaisons de moyennes, sans qu'aucun test de vraisemblance ne soit mis en œuvre<sup>5</sup>. Seul l'article de Sandrine Henry et Berthille Pallaud, qui esquissent un rapprochement intéressant entre amorces de mots et répétitions, fait un usage systématique de tests de pertinence même si les phénomènes d'hapax semblent négligés dans certaines discussions. Au total, il semblerait que les chercheurs du groupe DELIC se soient auto-persuadés de la représentativité absolue du corpus CRFP ! Ce n'est pas faire injure au travail déjà réalisé que de les rappeler à plus de correction méthodologique. Notons à leur décharge que ces insuffisances se retrouvent dans la plupart des recherches francophones en linguistique de corpus, alors même que c'est en France qu'est née la statistique lexicale...

Ces réserves méthodologiques ne doivent en rien troubler le lecteur à la fois sur l'importance du *Corpus de Référence du Français Parlé* et sur l'intérêt de ces études. Comme je l'ai déjà évoqué avec la contribution de Sandrine Candéo, ces travaux montrent combien l'analyse de faits réels en corpus peut conduire à remettre en cause certaines fausses intuitions. Les deux articles sur les structures détachées avec pronoms disjoints (Mylène Blasco-Dulbecco, de l'université de Clermont-Ferrand II et Sandrine Cadéo) montrent ainsi l'existence de structures fortement liées (*moi je* et *lui+SN*) qui doivent être considérées comme des constructions syntaxiques solidaires et indivisibles<sup>6</sup>. Bien que déjà relevé par certains auteurs, ce constat n'allait pas de soi *a priori*.

De même que la psychologie cognitive a su se départir des travers de l'analyse introspective, la linguistique de corpus est ainsi un garde-fou salutaire aux errements d'une recherche linguistique purement intuitive. De ce point de vue, ce dix-huitième volume des *Recherches sur le français parlé* représente une bonne illustration de l'apport des études sur corpus oraux. Le chercheur travaillant sur un des thèmes

---

<sup>5</sup> Dans son article sur les formes disjointes des pronoms sujets, Paul Cappeau évoque bien un test du Chi-Deux sur une observation particulière. Concernant les distributions avec les formes en *lui*, il ne cherche pourtant pas à étudier les différences observées sur les corpus *CRFP* et *CorpAix*. À première vue, celles-ci ne me semblent pas statistiquement négligeables.

<sup>6</sup> ou plus précisément comme une construction nominale avec extension apposée dans le second cas.

abordés y découvrira des résultats qui l'intéresseront même si ces recherches, de l'aveu même de leurs auteurs, doivent être étendues : le CRFP sort en effet tout juste des limbes !

Mais surtout, je ne peux qu'espérer que cet ouvrage rencontrera l'attention de linguistes s'interrogeant sur l'utilisation de corpus dans leurs recherches ou plus précisément sur l'apport du *Corpus de Référence du Français Parlé* par rapport aux ressources orales déjà existantes. Cet ouvrage mosaïque offre en effet un excellent panorama des possibilités offertes par le CRFP. Nul doute qu'il en existe d'autres, d'ailleurs...

## Sommaire de l'ouvrage

- Présentation, P. CAPPEAU,
- Présentation du *Corpus de Référence du français parlé*, Équipe DELIC
- Les compléments de lieu réalisés par y : description des usages F. SABIO
- Quelques données sur les adverbes en *ment* dans le corpus de référence de français parlé, M. BILGER
- Fréquence des marqueurs discursifs en français parlé : quelques problèmes de méthodologie, C. CHANET
- Les formes disjointes des pronoms sujets, P. CAPPEAU
- Quelques éclairages sur le sujet type *moi je* à l'oral, M. BLASCO-DULBECCO
- *Lui le propriétaire, le propriétaire, lui* : deux constructions bien distinctes, S. CADDEO
- Du bon usage des amorces dans la transcription des corpus, M.-N. ROUBAUD
- Étude des interactions entre pauses silencieuses et pauses remplies en français parlé E. CAMPIONE
- Amorces de mots et répétitions dans les énoncés oraux S. HENRY, B. PALLAUD

---

## L'auteur de la revue critique



Jean-Yves Antoine est professeur en Informatique à l'Université François Rabelais de Tours (laboratoire LI). Après un doctorat sur la compréhension de parole préparé à l'Institut de la Communication Parlée (Grenoble), il a mené des études post-doctorales sur le même thème au CLIPS-IMAG (Grenoble) avant de rejoindre le laboratoire VALORIA (Université de Bretagne Sud) en qualité de maître de conférences. Jusqu'à cette année, il y a dirigé le groupe CORAIL de recherche en ingénierie linguistique. Il conduit des travaux sur le dialogue homme-machine, l'aide à la communication pour personnes handicapées, l'évaluation des systèmes d'ingénierie des langues ainsi que des recherches en linguistique de corpus. Il anime un groupe de recherche du PRC-I3 sur la compréhension de la langue. Il est enfin rédacteur en chef des *Cahiers Romains de Sciences Cognitives*.

## La réaction des auteurs

J'approuve en tous points la critique fort pertinente de Jean-Yves Antoine. Le terme Corpus de Référence est manifestement mal choisi, puisque le CRFP ne correspond pas à la définition, largement acceptée, de Sinclair (1996) :

*"A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials."*

Le CRFP dans son état actuel ne peut prétendre à cette appellation ni dans sa taille, hélas trop faible malgré les efforts considérables que sa constitution a nécessités, ni par sa composition qui est totalement déséquilibrée en faveur d'un genre de parole particulier, celui de l'enquête linguistique (entretien sollicité), qui représente 81 % des enregistrements. Je ne crois pas à la représentativité, notion qui est souvent associée à celle de référence, car il faudrait préciser « représentativité de quoi » (pour reprendre l'expression de Habert, 2000) : y aurait-il un locuteur « moyen » dont on pourrait ainsi représenter le langage ? Étant donné l'immense variété individuelle, sociale, géographique et temporelle des usages langagiers, l'entreprise est vaine, faute d'une enquête sociolinguistique d'une ampleur impossible à envisager dans le futur prévisible. Je préfère remplacer cette notion par celle de *variété* : il serait souhaitable que nos corpus contiennent le plus de situations de parole possibles, à charge pour le linguiste de les utiliser avec des méthodologies adaptées.

Le point de vue de l'équipe a changé depuis le lancement du CRFP (1998) : il n'est plus considéré comme le corpus de référence du français parlé, mais comme une simple tranche d'un tel corpus, en construction. Nous l'avons d'ailleurs renommé CRFP-1, pour bien marquer ce changement de point de vue. Une deuxième tranche est d'ores et déjà en construction sous la direction d'André Valli, qui sera constituée de français « cultivé » des médias (émissions culturelles, débats, etc.). Par ailleurs, l'équipe a entrepris un programme de restauration des corpus du GARS, et de classement des corpus DELIC réalisés depuis 2000, dans la perspective d'un tel « corpus évolutif de référence » du français. Le travail de recensement entrepris par Estelle Campione, fait d'ores et déjà apparaître 3 millions de mots de transcriptions disponibles sous forme informatique. Un lent travail d'harmonisation des formats et des conventions de transcription est en cours, notamment grâce au travail de Marie-Noëlle Roubaud, qui s'attache à la restauration des corpus les plus anciens. Ce travail patient, et quasiment archéologique, fait apparaître une diversité plus importante que prévu des situations de parole, même si l'enquête-entretien est prédominante. Il n'en demeure pas moins que la remarque de Jean-Yves Antoine est fondée. Au total, les corpus GARS-DELIC contiennent peu de dialogues, ce qui s'explique à la fois par des raisons historiques et techniques : d'une part, les études du GARS, principalement de nature syntaxique, trouvaient un meilleur support dans des quasi-monologues que dans les dialogues et multilogues, dont la « fragmentation » et les aspects pragmatiques compliquaient trop l'étude en première approche ; d'autre part, l'enregistrement de conversations et multilogues réellement spontanés est difficile du point de vue technique (à moins de se restreindre aux dialogues téléphoniques, ce qui est loin de refléter la variabilité souhaitée). Cette

situation est en train de changer : le travail accumulé par l'équipe en analyse syntaxique permet désormais d'examiner des situations plus complexes, et l'évolution des techniques (micros FM, enregistrement digital, etc.) laisse envisager des conditions d'enregistrement de multilogues en conditions correctes « hors laboratoire ».

La deuxième critique de fond de Jean-Yves Antoine concerne les aspects quantitatifs des études linguistiques présentées dans le volume. Que peut-on ajouter à ce qui a été dit ? L'aspect quantitatif des études présentées est du même ordre que dans la plupart des travaux de « pure » linguistique, et se limite généralement à des dénombrements et des moyennes. Il ne faut sans doute les considérer que comme des indications heuristiques, un premier « débroussaillage » du terrain. C'est, je crois, non seulement l'équipe DELIC qui est concernée, mais toute la recherche en linguistique contemporaine, qui a bien du mal à adopter les méthodes scientifiques qui ont cours dans les autres sciences (y compris humaines, comme la psychologie expérimentale ou la sociologie). L'équipe est en train « d'absorber » un premier choc culturel, celui de l'informatisation des corpus : numériser les enregistrements, constituer des bases de données, transcrire avec des outils informatiques, utiliser des logiciels de traitement divers (étiquetage morpho-syntaxique, etc.), sont déjà des opérations qui ne vont pas nécessairement de soi dans le milieu de la linguistique traditionnelle. L'étape d'intégration suivante, celle consistant à adopter des méthodes quantitatives fouillées, me paraît plus lointaine. Il est vrai que la France peut s'enorgueillir de recherches de pointe en statistique linguistique (principalement lexicale), avec les travaux de Guiraud, Muller, Brunet et bien d'autres. Toutefois, les communautés restent bien cloisonnées, et ce savoir-faire a beaucoup de mal à perfer dans les milieux linguistiques traditionnels (les cursus linguistiques ne lui donnent d'ailleurs pratiquement aucune place). L'équipe DELIC essaie d'avancer modestement dans cette direction, mais le chemin ne peut qu'être long et difficile.

Enfin, une précision technique : le corpus est effectivement aligné avec le son au format XML, mais l'alignement est aussi accessible directement à partir du concordancier *Contextes* développé dans l'équipe<sup>7</sup>, qui permet au linguiste d'écouter les fragments sonores par un simple clic sur les passages de la transcription. Le son peut d'ailleurs être compressé au format WMA ou MP3, ce qui permet une qualité suffisante pour les utilisations courantes.

Jean Véronis, responsable équipe DELIC

## Références Bibliographiques

Habert B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? In Bilger M. (dir.), *Linguistique sur corpus. Études et réflexions*. Perpignan : Presses Universitaires de Perpignan. 11-58.

Sinclair J. (1996). EAGLES. Preliminary recommendations on Corpus Typology. EAG—TCWG—CTYP/P. Version of May, 1996. En ligne : <http://www.ilc.cnr.it/EAGLES96/corpusTyp/corpusTyp.html> (Dernière date de consultation : 24/10/2004).

---

<sup>7</sup> <http://www.up.univ-mrs.fr/veronis/logiciels/Contextes>