Mutualisation et uniformisation de ressources de français parlé

Christophe Benzitoun, Lolita Bérard

Christophe Benzitoun, ATILF CNRS & Université de Lorraine
Lolita Bérard, ATILF CNRS CNRS & Université de Lorraine

Mutualisation et uniformisation de ressources de français parlé

Résumé

La constitution de corpus oraux étant très coûteuse, il apparaît à l'heure actuelle incontournable de mettre des ressources en commun afin d'obtenir un corpus significatif pour la description du français. Notre expérience a consisté à réunir six corpus de français parlé, chacun étant transcrit et annoté en fonction d'un objectif particulier (sociolinguistique, phonologique, syntaxique). Nous présentons dans cet article la méthodologie adoptée pour obtenir des données unifiées et mises en forme dans une optique d'utilisation de logiciels d'enrichissement/exploitation de corpus. En effet, l'étape suivante sera de recourir à des outils de traitement automatique de corpus (analyseurs morpho-syntaxiques) ainsi que des outils de requêtes (concordanciers). Cette expérience nous a montré à quel point il est primordial de définir des recommandations pour les conventions de transcription, afin de faciliter l'échange et la mise en commun des données.

Abstract

According to the cost of speech transcription, it is very important to pool data to obtain a big size corpus to describe French. Our work consisted to pool six spoken French corpora, each with a specific goal (sociolinguistics, phonology, syntax), to format them for automatic exploitations. Indeed the next step will be to use NLP corpus tools (tagger, parser, concordancer). This experience showed that it is very important to specify recommendations for transcription conventions to make easier sharing and pooling data.

Mots-clés : corpus oraux, mutualisation, transcription, informatisation
Keywords: speech corpora, pooling, transcription, computation

Contexte

Depuis de nombreuses années, les linguistes s'accordent à dire que le français souffre d'un retard important dans l'élaboration de corpus, notamment oraux (Bilger 2000a, 2000b, Bruxelles et alii 2009). Il existe en fait de nombreuses transcriptions mais celles-ci sont dispersées, hétérogènes et majoritairement peu accessibles (Cappeau & Seijido, 2005). Or, dans cette pénurie et surtout cet éparpillement généralisé, il semble important de savoir dans quelle mesure ces données sont mutualisables. En effet, dans le contexte actuel, c'est sans doute la seule solution envisageable si l'on souhaite un jour doter le français d'un corpus de taille comparable au *British National Corpus*. En l'absence d'un tel « corpus de référence », considéré comme « représentatif » de la langue, il est extrêmement difficile de produire des études systématiques sur le français contemporain basées sur corpus, notamment des études lexicales, grammaticales et syntaxiques. Et à défaut d'un corpus de référence, notion sans doute difficile à cerner de manière précise, un corpus d'au moins 5-6 millions de mots présentant des situations variées représenterait déjà une base de travail raisonnable. Pour l'heure, les différents corpus adoptent des protocoles de constitution, de transcription et d'annotation extrêmement variés. Il n'est donc pas possible de se servir de ces données directement. Une simple concaténation mènerait certainement à des résultats erronés, les motifs recherchés ayant des formes potentiellement hétérogènes et difficilement prédictibles.

Nous nous sommes donc attachés à uniformiser le plus possible les transcriptions auxquelles nous avons eu accès (en faisant attention à ne pas trop les dénaturer), dans la perspective de constituer une archive orale de langue française importante, unifiée et compatible avec des outils informatiques. Bien évidemment, nous sommes pleinement conscients du fait que nous avons conçu cette mutualisation en

fonction de nos objectifs (analyses syntaxique, lexicale et grammaticale) et que, par là même, ces données ne seront pas exploitables pour certaines visées, interactionnelles notamment. Mais comme l'ont souligné notamment Cappeau et Gadet (2007), la seule récupération, par d'autres chercheurs, de données antérieures pose déjà des problèmes pour mener de nouvelles études. De plus, la réduction des transcriptions au plus petit dénominateur commun en termes de conventions de transcription hétérogènes rend totalement illusoire l'exploitation tout azimut. Mais signalons toutefois que les premières exploitations que nous avons menées sont très encourageantes (cf. Benzitoun et alii, 2010).
En outre, notre objectif est d'exploiter ce corpus à l'aide de logiciels (concordanciers, traitements lexicométriques, étiquetage morphosyntaxique, etc.), la quantité de données recueillie rendant une lecture systématique totalement inenvisageable. Notre travail va donc au-delà de celui du Centre de Ressources pour la Description de l'Oral (CRDO), dont l'objectif d'unification des conventions de transcription était moins prononcé et les modifications des fichiers très limitées.
Dans le cadre de notre projet, nous avons regroupé les corpus suivants (les références des corpus se trouvent après la bibliographie) :

- Corpus de Français Parlé Parisien [CFPP] : 24 enregistrements, 350 000 mots
- C-ORAL-ROM (uniquement la partie en français) : 175 enregistrements, 300 000 mots
- Corpaix : 265 transcriptions, malheureusement sans le son ni les métadonnées, 1 000 000 mots
- Corpus de Référence du Français Parlé [CRFP] : 134 enregistrements, 435 000 mots
- Choix de textes de français parlé [CTFP] : 36 extraits enregistrés entre 1989 et 2000, 70 000 mots
- Phonologie du Français Contemporain (uniquement les discussions libres enregistrées en France) [PFC] : 163 enregistrements, 250 000 mots

Ces corpus sont transcrits intégralement en orthographe standard, ou avec quelques cas très restreints mettant en jeu des artifices orthographiques, et ils ne sont pas surchargés de symboles additifs « para-orthographiques » (intonation, ponctuation, etc.), même s'ils n'en sont pas dépourvus pour autant. En outre, ils possèdent des enregistrements effectués dans des situations très variées (conversations, entretiens, réunions, émissions de radio et de télévision, témoignages, parodies, recettes, messages laissés sur un répondeur, cours, souvenirs, récits de vie, situations formelles et informelles, locuteurs enfants et adultes) avec une prépondérance de l'oral spontané, ou en tout cas non préparé. Ce type de production nous intéresse tout particulièrement, car, à la suite notamment de Sinclair et de Cappeau, nous envisageons l'oral de la manière suivante :

> *Many language scholars and teachers believe that the spoken form of the language is a better guide to the fundamental organization of the language than the written form.* (Sinclair, 1991, p.15)

> *On a supposé aussi que [...] avec l'oral on accédait aux faits majeurs de la distribution, qui était ainsi moins "parasitée" par des phénomènes d'ordre stylistique (et donc plus atypiques).* (Cappeau, 2002, p.11)

Et nous émettons l'hypothèse que le français non planifié est un support privilégié dans lequel les phénomènes d'organisation fondamentale de la langue sont les plus visibles.
Parallèlement à ce travail d'unification, nous avons fait un bilan d'une partie des pratiques actuelles observées dans le domaine de la constitution de corpus de français parlé. Ce bilan montre l'importance fondamentale de proposer des solutions dans le cadre par exemple de la TEI (*Text Encoding Initiative*), en tenant compte à la fois de la dimension linguistique et du besoin de normalisation inhérent à l'informatisation. Nos interventions sur les transcriptions mettent d'ailleurs en évidence la prise en compte encore trop superficielle des contraintes liées à l'informatisation, parfois sous-estimées par les concepteurs de corpus. Or, à l'heure actuelle, des guides de bonnes pratiques numériques[1] voient le jour, ce qui devrait permettre d'améliorer de manière significative l'informatisation des corpus oraux.

[1] On peut signaler par exemple les adresses suivantes : http://www.tge-adonis.fr/ressources/guides et http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm.

Protocole de mutualisation

Pour rendre les données exploitables une fois réunies, il est indispensable d'uniformiser les transcriptions et les symboles : il y a d'une part des symboles identiques qui ont des significations différentes suivant les corpus et d'autre part des symboles différents qui renvoient aux mêmes phénomènes. Il est également indispensable que tous ces corpus soient disponibles dans le même format. Nous avons choisi comme format de base celui utilisé par le logiciel Transcriber[2], qui comporte un balisage XML. A partir de ce format, nous avons développé des outils de conversion vers des formats compatibles avec certains logiciels d'interrogation.
Voici les traitements que nous avons effectués.

Doublons

Les transcriptions de certains enregistrements peuvent se retrouver dans plusieurs corpus. Les conventions étant généralement différentes, il paraissait périlleux d'effectuer un repérage automatique des doublons. Nous avons donc procédé à des sondages (recherche d'un terme du début et de la fin de chaque transcription) et à la lecture de certains passages pour vérification. Nous avons rencontré différents cas de figure : copie totale, partielle, et même, transcriptions remaniées. Nous avons supprimé les fichiers dont tout le contenu était copié. Lorsqu'une partie seulement était commune, nous avons conservé les deux transcriptions en supprimant cette partie commune dans l'un des deux fichiers. Si un doublon à supprimer était mieux transcrit, nous l'avons injecté en lieu et place du passage correspondant dans la transcription initiale. Cela permettait simplement de réduire les corrections, les conventions d'annotation étant de toute façon uniformisées par la suite. Avec ce traitement, nous sommes passés de 826 à 749 transcriptions. Ces dernières représentent environ 2.300.000 mots. Sur cette masse, nous possédons environ cent trente heures d'enregistrements audio.

Uniformisation des transcriptions

Afin de ne pas rencontrer de problèmes lors de la phase d'exploitation, nous avons essayé de limiter les caractères à ceux présents dans la table ASCII étendu. Hormis les caractères parasites, dus vraisemblablement à une ancienne conversion du logiciel de traitement de texte *Microsoft Word* vers un format texte brut ou à un système de reconnaissance automatique de caractères, les modifications que nous avons faites étaient liées à la minutie et aux habitudes des transcripteurs. Il pouvait par exemple y avoir des majuscules accentuées ainsi que des lettres collées 'œ'. De nombreux logiciels gèrent mal ces cas de figure et il arrive couramment que les personnes qui constituent les données oublient ces paramètres. Ces caractères spéciaux ont initialement été repérés au cours des autres phases du traitement. En cours de route, nous avons utilisé *WordSmithTools*[3] afin de dresser la liste des caractères présents dans les fichiers, ce qui a beaucoup accéléré le travail et l'a rendu plus fiable.
Le problème le plus complexe, mais également le plus intéressant d'un point de vue linguistique, réside dans l'étude des différentes conventions de transcription choisies ainsi que dans le choix d'en adopter une plutôt qu'une autre. Dans ce domaine, les compromis avec les paramètres techniques sont incontournables. Comme nous l'avons dit, tous les corpus que nous avons traités utilisent l'orthographe standard (cf. Blanche-Benveniste et Jeanjean, 1987). Mais il y a cependant quelques exceptions tels *d'la, t'as, y'a*. Selon nous, ces conventions posent au moins deux problèmes. Le premier, c'est que les deux versions existent dans les corpus (*de la / d'la*). Il est donc vraisemblable de penser que les utilisateurs ne penseront pas à rechercher les deux formes et le résultat de leur requête sera biaisé. Le deuxième, c'est que les programmes d'annotation automatique (parseurs, taggueurs, etc.) ont généralement été élaborés pour l'écrit et n'attendent pas des suites telles que celles-ci. En attendant que soient disponibles des outils adaptés à l'oral et aux trucages orthographiques, nous avons décidé de tout basculer en orthographe standard.

Uniformisation des annotations

Nous avons choisi de supprimer toutes les marques de ponctuation. Ne pas disposer d'un texte ponctué peut poser des problèmes à certains lecteurs, mais la plupart des transcriptions n'étant pas ponctuées, il était plus simple de supprimer des informations (les ajouter aurait demandé un travail extrêmement

[2] http://trans.sourceforge.net.
[3] http://www.lexically.net/wordsmith/.

long). De plus, il a été souligné que la ponctuation de l'écrit n'était pas forcément adaptée à l'oral et que cela imposait une analyse préalable (Blanche-Benveniste et Jeanjean, 1987). Il est donc gênant d'avoir des transcriptions ponctuées alors qu'elles se veulent être à la base des analyses ultérieures. Nous avons également supprimé les majuscules suivant ces ponctuations.
De même, nous avons supprimé toutes les marques d'intonations, de pauses et d'allongements. Ce choix peut paraître trop radical et il aurait sans doute été possible de conserver l'intégralité des informations marquées à l'aide de symboles sous forme de balises, puis de les supprimer et de les réinjecter à volonté. Mais nos compétences en informatique rendaient les résultats peu sûrs et ces informations n'étaient pas directement utiles pour nous. De plus, cela aurait eu l'inconvénient de gêner certains logiciels. En outre, nous avons remarqué que l'annotation de ces phénomènes n'était pas systématique, car elle est liée à la subjectivité de chaque transcripteur.
Certains phénomènes étaient marqués uniquement dans un ou deux corpus : le discours rapporté (par des guillemets simples ou doubles), les répétitions *(&le &le*), les pauses remplies (*&euh*). Pour faciliter l'uniformisation et ne pas surcharger les transcriptions, nous avons choisi de supprimer également ces annotations (guillemets et &), en conservant bien sûr l'intégralité de ce qui avait été prononcé (*le le* ; *euh*).
Pour le reste, nous avons conservé et étendu les conventions les plus fréquentes (multi-transcription = /ces, ses/, amorce = am-, suite de phonèmes incompréhensibles = XXX). Les commentaires (rires, bruits ambiants et événements ponctuels) ont été placés dans les balises proposées par défaut par le logiciel *Transcriber* (balises <Comment> ou <Event>) car ils ne correspondent pas à des paroles effectivement prononcées par les locuteurs. Ainsi, si l'on recherchait « rire(s) » dans les corpus originaux, on obtenait 1980 résultats alors qu'après le balisage, on obtient seulement 69 occurrences. Il était auparavant impossible de distinguer de manière systématique les paroles prononcées par les locuteurs des commentaires du transcripteur et cela est évidemment très gênant quand on décrit la langue. De plus, l'utilisation du langage XML pour annoter ce type d'informations étant fort répandue à l'heure actuelle, il nous a semblé indispensable de mettre dans ce langage tout ce qui pouvait l'être.
Nous avons également rencontré deux types de problèmes qui concernent les chevauchements de parole. Il arrive que les transcripteurs, dans le but d'indiquer scrupuleusement à quel endroit le chevauchement commence, aient coupé un mot en deux. Cette manière de procéder présente l'inconvénient de ne pas pouvoir repérer un mot à l'aide d'outils utilisant les chaînes de caractères. Les mots coupés qui ont été repérés ont alors été recollés. De même, à l'intérieur d'un chevauchement, les paroles d'un même locuteur peuvent être fragmentées en plusieurs tours. Lorsque le chevauchement avait lieu entre un locuteur qui produisait un énoncé et un autre qui acquiesçait de temps à autre, nous avons supprimé les acquiescements et rassemblé l'énoncé, bien que cela soit problématique pour l'alignement automatique et l'analyse de discours. Néanmoins, cela est utile pour récupérer des constructions syntaxiques et leurs contextes ; nous envisageons même de coller tous les tours de parole d'un même locuteur, afin de réunir les syntagmes, comme la relative (*qui a deux enfants*) et sa tête (*Anne*) dans le passage suivant, extrait de PFC (pour des raisons de lisibilité, nous avons aménagé le format dans le passage cité) :

L2 j'ai que Anne
L1 tu as
L2 j'ai que Anne
L1 que Anne
L2 qui a deux enfants
L1 deux enfants
(11aml1)

Résultat souhaité :

L2 j'ai que Anne
L1 tu as que Anne
L2 j'ai que Anne qui a deux enfants
L1 deux enfants

La majorité des traitements n'a pas pu être effectuée de manière automatique (certaines marques étant identiques pour identifier plusieurs phénomènes dans un même corpus ou entre corpus). Et même lorsque l'automatisation était possible, une longue vérification était indispensable. En effet, il existe de nombreuses variations, indépendamment des conventions fournies aux transcripteurs. Pour les multi-transcriptions, par exemple, il existe un grand nombre de combinaisons possibles :

- marques de début et de fin : parenthèse, crochet, slash [celle de fin n'étant pas forcément identique à celle du début] ;
- marque de séparation des deux formes : virgule, point virgule ;
- présence ou non d'espace(s) avant et après chaque marque.

Conversion vers un format de base : Transcriber

Au départ, les transcriptions avaient trois formats distincts : texte brut (ASCII), *Transcriber* et *Praat*. Nous avons choisi comme format de sortie le format *Transcriber*, pour convertir le moins de fichiers possible (une partie non négligeable des transcriptions étant déjà dans ce format). Pour cela, nous avons ajouté des balises à chaque tour de parole (<Turn> pour symboliser les tours de parole et <Sync> pour la synchronisation avec le son) et nous avons renseigné pour chacun les attributs « startTime » (début du segment sonore), « endTime » (fin du segment sonore) et « speaker » (locuteur), ainsi que la balise <Who> (locuteur) en cas de chevauchements de parole.
Certains corpus ont nécessité un traitement plus conséquent pour obtenir la conversion souhaitée. Pour C-ORAL-ROM, par exemple, nous disposions d'un côté d'un fichier d'alignement texte-son et de l'autre de la transcription orthographique. Il a donc fallu élaborer un programme en *Python* afin de convertir le corpus en format *Transcriber* valide[4]. La conversion n'a pas été nécessaire pour Corpaix qui ne dispose pas d'alignement texte-son.

Projection vers les formats d'entrée de logiciels

Une fois l'uniformisation des corpus effectuée et vérifiée, nous avons entamé la projection vers le format d'entrée de certains logiciels d'exploitation. Pour l'instant, nous avons effectué la conversion pour le concordancier *Contextes* et mené des tests avec les outils d'analyse textométriques *Le Trameur* et *Lexico 3*. Avec *Contextes*, l'alignement texte-son est directement exploitable par le logiciel et l'utilisateur peut écouter le segment sonore qui l'intéresse. Nous disposons également de la mention du locuteur courant, ce qui permet de dissocier les paroles prononcées par des locuteurs différents.

Un exemple

Afin d'illustrer précisément en quoi ont consisté nos interventions sur les transcriptions initiales, voici un exemple de transformation tiré de [PFC, 44ajs1lg]. Les éléments surlignés correspondent à ce que nous avons modifié (et éventuellement ajouté ou supprimé).

JS : Et puis voilà com/ pourquoi on est parti là-bas. <E : D'accord.>
JS : Euh, ç'aurait été, euh, des prix plus abordables pour nous, vraisemblablement on, on aurait été à La Baule, mais euh, pour nous c'était déjà beaucoup trop cher.<E : (XX).>
E : Oui, puisque déjà à lépoque, euh. <JS : Ah bah on payait, j'ai payé mon terrain vingt francs le mètre.>

Figure 1 : Transcription originale

```
<Turn speaker="spk2 spk1" startTime="81.9641076057252" endTime="84.97338973793642">
<Sync time="81.9641076057252"/>
<Who nb="1"/>
et puis voilà com- pourquoi on est partis là-bas
<Who nb="2"/>
d'accord
</Turn>
<Turn speaker="spk2 spk1" startTime="84.97338973793642" endTime="91.727725794005806">
<Sync time="84.97338973793642"/>
<Who nb="1"/>
euh ç'aurait été euh des prix plus abordables pour nous vraisemblablement on on aurait été à La Baule mais euh pour nous c'était déjà beaucoup trop cher
<Who nb="2"/>
XX
</Turn>
```

[4] Nous remercions B. GAIFFE pour avoir effectué ce programme de conversion.

```
<Turn speaker="spk1 spk2" startTime="91.7277257940058" endTime="96.32603510128133">
<Sync time="91.7277257940058"/>
<Who nb="1"/>
oui puisque déjà à l'époque euh
<Who nb="2"/>
ah bah on payait j'ai payé mon terrain vingt francs le mètre
</Turn>
```

Figure 2 : Transcription finale après transformation

Bilan et perspectives

Dans ce qui précède, nous nous sommes attachés à décrire la méthodologie choisie pour uniformiser plusieurs corpus, sans forcément mettre l'accent sur les aspects les plus problématiques. Or, bien sûr, la mutualisation des données et les choix effectués amènent à la perte de certaines informations et rendent ainsi les données exploitables seulement pour des recherches en nombre limité. Mais il est nécessaire de normaliser les données si l'on veut pouvoir les utiliser de manière efficace et formuler des conclusions sur l'utilisation de la langue en se basant sur de grands corpus. De plus, les exploitations que nous envisageons étant essentiellement lexico-grammaticales, cette perspective a évidemment eu une incidence sur nos choix. Toutes les modifications ont été enregistrées dans une seconde version des transcriptions, afin de ne pas dénaturer les corpus initiaux. Nous avons également élaboré une première version normalisée respectant les conventions propres à chaque corpus, en essayant de corriger les passages où les transcripteurs ne les avaient pas respectées.

Nos traitements ont été documentés de manière systématique afin de ne pas risquer que certaines étapes se télescopent ou génèrent des erreurs et afin que l'on sache exactement sous quelle forme se présente la banque de données finale. Toutefois, il serait intéressant, à partir de notre travail de synthèse des différentes conventions de transcription, de faire des propositions d'encodage en TEI afin d'aboutir à des pratiques diffusables et ainsi ne pas répéter à l'infini le travail que nous avons effectué. Mais pour ce faire, il faut que la communauté francophone des corpus oraux s'empare de la question et développe un cahier des bonnes pratiques numériques, dans la lignée de Baude (2006).

La base actuelle s'enrichit en permanence grâce au projet *Traitement des Corpus Oraux du Français* (TCOF[5]) qui est mené au sein du laboratoire ATILF et qui est très proche du format que nous avons adopté. Nous pourrions aussi intégrer le corpus Parole Publique de l'université de Tours, qui possède des conventions proches. D'ici peu, nous devrions donc atteindre la barre des 3.000.000 de mots de français parlé.

La question de la diffusion est évidemment épineuse. Etant donné que les transcriptions proviennent de différentes sources et donc ont été menées par divers responsables, ce sont eux qui peuvent décider de leur rediffusion. Nous comptons nous renseigner pour connaître leur avis quant à la possibilité de diffuser les transcriptions dans notre format et ce, dans les conditions qu'ils fixeront. D'ores et déjà nous savons que certains corpus sont protégés et qu'une diffusion libre de la totalité des données n'est pas envisageable. En outre, cette rediffusion ne serait intéressante que dans l'optique d'une possible interrogation de l'ensemble des transcriptions. Pour ce faire, un logiciel libre, comparable à Contextes, est en cours de développement à l'ATILF en coopération avec le LORIA et une version bêta a déjà vu le jour.

Les métadonnées à notre disposition ont été projetées sous un même format, à savoir un format tabulaire, afin de les rendre interrogeables. La phase de normalisation de tous les jeux de métadonnées est en cours afin de rassembler sous une même étiquette les informations identiques ou très proches. Il faudra ensuite développer un système de requêtes portant sur les champs afin que chacun puisse créer ses propres sous-corpus.

Nous avons déjà eu l'occasion d'utiliser ces transcriptions dans le cadre de travaux de description sur *prochain* (Benzitoun et alii, 2010). Ainsi, nous avons pu constater que la base est opérationnelle et qu'elle permet de décrire des phénomènes impossibles à décrire avec un corpus de taille moindre (pour

[5] Les transcriptions du projet TCOF sont librement téléchargeables à l'adresse : http://www.cnrtl.fr/corpus/tcof/. Des transcriptions supplémentaires seront ajoutées au fur et à mesure qu'elles seront finalisées.

prochain, nous avons relevé seulement 164 occurrences dans tout le corpus). Et malgré l'absence d'un objectif initial commun des différents corpus que nous avons réunis, nous pensons qu'il est tout de même possible de nous en servir, eu égard à ces premiers résultats. Il faut malgré tout rester très prudent et être particulièrement vigilant quant aux éventuels effets de bord liés à des surreprésentations d'un type de parole particulier.

Références bibliographiques

BAUDE O. (coord.), 2006, *Corpus oraux : Guide des bonnes pratiques*. Presses universitaires d'Orléans, CNRS Editions.

BENZITOUN C., BRESSON S., BUDZINSKI L., DEBAISIEUX J.-M. & HOLZHEIMER K., 2010, « Quand un corpus rencontre un adjectif du troisième type. Etude distributionnelle de *prochain* », M. OLIVIERI (éd.), La syntaxe de corpus. *Corpus 9*, 245-264.

BILGER M., 2000a, *Corpus. Méthodologie et applications linguistiques*. Paris, Honoré Champion et Presses Universitaires de Perpignan.

BILGER M., 2000b, « Linguistique sur corpus : étude et réflexion », *Cahier de l'Université de Perpignan 31*, Perpignan, Presse universitaire de Perpignan.

BLANCHE-BENVENISTE C. & JEANJEAN C., 1987, *Le français parlé : transcription et édition*. Didier érudition.

BRUXELLES S., MONDADA L., TRAVERSO V. & SIMON A.C. (éd.), 2009, « Grands corpus de français parlé : bilan historique et perspectives de recherches », *Cahiers de linguistique* 33(2)

CAPPEAU P., 2002, « Entre l'auxiliaire et le participe passé », *Recherches sur le français parlé 17*, 11-28.

CAPPEAU P. & SEIJIDO M., 2005, « Les corpus oraux en français (inventaire 2005 v1.0) », DGLFLF. http://www.culture.gouv.fr/culture/dglf/recherche/corpus_parole/Presentation_Inventaire.pdf.

CAPPEAU P. et GADET F., 2007, « L'exploitation sociolinguistique des grands corpus, Maître-mot et pierre philosophale », *Revue Française de Linguistique Appliquée*, *12*(1), 99-110.

SINCLAIR J., 1991, *Corpus Concordance Collocation*. Oxford, Oxford University Press.

Corpus utilisés ou cités

Choix de Textes de Français Parlé (CTFP) :

BLANCHE-BENVENISTE C., ROUGET C. & SABIO F., 2002, *Choix de textes de Français parlé, 36 extraits*. Paris, Champion, Collection Les français parlés.

C-Oral-Rom : http://lablita.dit.unifi.it/coralrom/.

CRESTI E. & MONEGLIA M., 2005, « Integrated Reference Corpora for Spoken Romance Languages » *Studies in corpus linguistics 15*, John Benjamins.

CorpAix :

BLANCHE-BENVENISTE C., 1999, « Constitution et utilisation d'un grand corpus, Grands corpus : diversité des objectifs, variété des approches », *Revue Française de Linguistique Appliquée 4*, 1.

Corpus de Français Parlé Parisien (CFPP) :

BRANCA-ROSOFF S., FLEURY S., LEFEUVRE F. & PIRES M.

Discours sur la ville, Corpus de Français Parlé Parisien des années 2000 (CFPP2000). http://ed268.univ-paris3.fr/CFPP2000/ ; corpus téléchargé en juin 2010

Corpus de Référence du Français Parlé (CRFP) :
Equipe DELIC, 2004, « Présentation du 'Corpus de référence du français parlé' », *Recherches sur le français parlé 18*, 11-42. http://sites.univ-provence.fr/veronis/pdf/2004-presentation-crfp.pdf.

Phonologie du Français Contemporain (PFC) :
http://www.projet-pfc.net/, corpus téléchargé en mars 2010
DURAND J., LAKS B. & LYCHE C.,
2002, « La phonologie du français contemporain : usages, variétés et structure », C. PUSCH & W. RAIBLE (eds.) *Romanistische Korpuslinguistik-Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language*. Tübingen, Gunter Narr Verlag, 93-106.
DURAND J., LAKS B. & LYCHE C.,
2005, « Un corpus numérisé pour la phonologie du français », G. WILLIAMS (éd.) *La linguistique de corpus*. Rennes, Presses Universitaires de Rennes, 205-217, Actes du colloque 'La linguistique de corpus', Lorient, 12-14 septembre 2002.

Corpus PAROLE PUBLIQUE de l'Université de Tours :
http://www.info.univ-tours.fr/~antoine/parole_publique/