Verbum XXX, 2008, 4

ET SI DEMAIN ON VOULAIT ÉLABORER UNE GRAMMAIRE DU FRANÇAIS PARLÉ SUR CORPUS… LA QUESTION DES DONNÉES

Christophe BENZITOUN
Université Nancy 2 & ATILF

RÉSUMÉ

Dans cet article, nous explorons la question de la faisabilité d'une "grammaire" du français parlé sur corpus à relativement court terme. Nous pointons le paradoxe que constituent les progrès de l'informatique et la rareté des études globales sur corpus oraux et nous avançons l'idée qu'il est utile d'élaborer le corpus d'étude simultanément au travail de description.

ABSTRACT

In this paper, we try to know if it is possible to make a "Grammar" of Spoken French during next years. A contemporary paradox is that there are not enough studies and corpora of Spoken French despite informatics progress. To go beyond this paradox, we suggest that it is useful to make corpora and descriptions at the same time.

1. OBJECTIFS DE LA "GRAMMAIRE"

Cela fait quelques années que périodiquement un projet collectif de grammaire sur corpus oraux refait surface au sein d'une communauté de chercheurs à laquelle j'appartiens. Mais, pour de multiples raisons, ce projet n'a pas encore pu se matérialiser. Or, compte tenu de l'importance que revêt un tel projet pour la société, j'ai décidé de procéder à un examen précis de sa faisabilité, en commençant par les données.

Nous avons choisi de dénommer ce projet "grammaire" faute de mieux, conscients que ce terme n'était pas vraiment adapté. Par exemple, une partie conséquente sera consacrée à la description des lexèmes en prenant en compte les constructions dans lesquelles ils s'insèrent. Nous étudie-

rons donc les environnements syntaxiques en tenant compte des lexicalisations et des collocations. La fréquence des tournures occupera également une place importante afin d'illustrer leur degré de représentativité.

Une autre originalité de notre projet réside dans l'emploi systématique de données authentiques comme sources de nos analyses et non l'inverse (approche dite "corpus driven"). La plupart des dictionnaires et des grammaires de langue française présentent des exemples attestés mais, à part quelques très rares exceptions[1], ils viennent essentiellement illustrer des classifications largement préexistantes. De plus, dans la tradition grammaticale de langue française, les exemples sont très souvent littéraires et très rarement oraux. Nous nous retrouvons finalement plus dans une approche anglo-saxonne illustrée par le *Collins Cobuild English Dictionary* et la *Longman Grammar of Spoken and Written English*.

Pour l'instant, notre projet est seulement au stade d'ébauche. En parallèle de la réflexion sur les données, nous faisons des tests à partir des formes *contre* et *cause(r)* pour esquisser deux entrées-types et proposer une démarche méthodologique. Cette mise en chantier est selon nous urgente car, même approximative, une description du type que nous envisageons est indispensable pour mieux connaître le français parlé aujourd'hui en France. Cela sera utile pour quantité d'études, de méthodes d'apprentissage du français ou bien encore comme base de comparaison pour détecter des pathologies, car comme l'a écrit Martinet (1974, 14, in D. François) :

> Ce n'est pas d'hier que les spécialistes des troubles du langage se plaignent de ne pouvoir se prononcer avec quelque assurance en l'absence d'une norme du français oral.

Le présent article est consacré à notre réflexion au sujet des données requises pour un tel projet. Plus précisément, il s'agira de dresser un état des lieux des contraintes, mais aussi de dire si notre grammaire peut être entamée dans un avenir relativement proche.

Pour ce faire, dans un premier temps, nous présenterons le travail qu'a nécessité l'élaboration du *Français Fondamental*. Puis nous ferons état des données disponibles et de leur éventuel "enrichissement". Avant de conclure en détaillant les données présentent dans notre archive.

2. LES PRÉCURSEURS

De nos jours, et même dans un passé relativement récent, aucune entreprise de description d'ensemble du français parlé n'a été lancée[2]. Or, en des temps plus anciens, quelques ouvrages ont vu le jour s'attachant à rendre

---

1 Pour l'écrit, on peut citer par exemple le dictionnaire *Le Trésor de la Langue Française* et la grammaire de Sandfeld (1936). Quelques autres exemples portant sur l'oral seront cités par la suite.

2 Il n'est qu'à lire la partie qu'y consacre Blanche-Benveniste dans le dernier tome de la monumentale *Histoire de la langue française* dédié aux années 1945-2000 pour s'en convaincre.

compte, principalement ou subsidiairement, du français parlé de l'époque. On peut citer, par exemple :

- Damourette & Pichon (1911-1946), qui citent plusieurs centaines de locuteurs ;
- Martinon (1927), lequel avait une visée plutôt puriste ;
- Bauche (1928), qui s'appuyait sur des exemples saisis à la volée, sans pouvoir prétendre à l'exhaustivité :

> […] j'ai simplement reproduit des phrases que j'ai entendues dans la rue, dans l'armée, dans les ateliers, les usines et les boutiques, chez les marchands de vin, dans les compartiments de troisième classe, dans les quartiers populaires de Paris, et, aussi, des phrases que j'ai collectionnées dans des lettres écrites par des gens du peuple. (Bauche, 1928, 29)

Dans une période plus récente, on peut citer seulement des travaux plus ciblés, tant au niveau des descriptions que des données, tels le *Français Fondamental* (années 50), s'intéressant essentiellement au vocabulaire, ou l'*Enquête Sociolinguistique à Orléans* (années 60) qui a donné lieu à l'élaboration d'un vaste recueil d'enregistrements. On peut mentionner également François (1974), qui expose un important travail d'analyse mais sur des données assez limitées, et les travaux de Blanche-Benveniste et Gadet (de nos jours), qui ont notamment mené un très grand nombre d'études de détails.

Dans la liste parcellaire que nous venons de présenter, il est frappant de voir que les ouvrages grammaticaux les plus généraux ont été réalisés il y a soixante ans (ou plus). Pourtant, les avancées technologiques récentes (essentiellement dans le domaine de l'informatique) ont grandement simplifié la conservation, la collecte et l'exploitation des données orales. Quand on fait état, par exemple, du travail laborieux qu'a nécessité le *Français Fondamental*, on ne peut être qu'étonné par l'inexistence de travaux comparables plus récents.

2.1. Les contraintes techniques du *Français Fondamental*

Le *Français Fondamental* est un lexique basé sur des transcriptions de français parlé. Il tient compte de la fréquence de chaque lemme et de leur répartition dans les différents enregistrements. Il aborde également quelques points de grammaire.

Au niveau des données sources, le *Français Fondamental* est composé de quatre enregistrements récupérés au Musée de la parole (français qualifié de "commun"), de neuf enregistrements provenant des Archives de la Radiodiffusion, de deux enregistrements appartenant aux Collections du Musée des Arts Populaires et d'enregistrements qu'ils ont réalisés eux-mêmes. À l'époque, la prise de son et la transcription étaient loin d'être aussi aisées qu'aujourd'hui, du moins du point de vue technique.

Les enquêteurs disposaient d'appareils à disques de papier magnétique "facilement transportables" dont le poids "peu élevé" faisait tout de même six kilos l'unité[3]. La prise de son nécessitait de changer de disques toutes les

3 Quand on les compare aux dictaphones numériques actuels, cela laisse assez songeur.

six minutes. Les disques ne pouvaient pas être conservés : ils étaient soit réutilisés soit effacés. Le simple contact accidentel des faces magnétisées aboutissait immanquablement à la destruction des enregistrements. Les enquêteurs disposaient également de magnétophones à bandes. Au final, les enregistrements de 275 témoins ont été retenus (sans échantillonnage particulier), ce qui correspond à 163 transcriptions différentes soit 312.135 mots.

Pour l'exploitation, les linguistes ont travaillé à partir de cahiers à feuilles mobiles qui étaient affectés à chaque texte. Les formes phonétiquement différentes des mêmes lemmes étaient notées séparément (*cheval/ chevaux*). Les formes verbales étaient considérées de la façon suivante :

AIMER 8 : aimer 2 ; aime (j') 2 ; aimait (il) 4

c'est-à-dire la forme-type (lemme) accompagnée du total des emplois de ce verbe et suivie de la liste de ses diverses formes conjuguées.

Pour obtenir la liste des fréquences, il suffisait alors de scinder les cahiers : toutes les pages 1 des 163 cahiers étaient réunies en un fascicule, les pages 2 également, et ainsi de suite… Au total, les auteurs ont retenu 7.995 formes différentes, ce qui a demandé le dépouillement de plusieurs centaines de milliers de feuillets.

2.2. Un *Français Fondamental* aujourd'hui ?

Au terme de cette rapide présentation de l'élaboration du *Français Fondamental*, on peut dire que la collecte des données orales ainsi que leur exploitation ont profondément été facilitées par la technique. Aujourd'hui, les dictaphones numériques permettent d'enregistrer en une fois plusieurs heures et de les transférer sur un ordinateur en seulement quelques minutes, le tout pour un prix raisonnable. De plus, l'extraction et le calcul de la fréquence des vocables sont réalisables automatiquement en quelques jours. Pourtant une entreprise similaire au *Français Fondamental* n'a même pas été lancée. Nous sommes en droit de nous demander pourquoi. N'y a-t-il aucun intérêt à comparer les résultats à cinquante ans d'intervalle ? Existe-il des problèmes de familiarisation des linguistes avec les outils informatisés, de disponibilité de ces outils ou des données ? Les données disponibles ne sont-elles pas suffisamment "représentatives" ? Pour obtenir des résultats rapidement, il faudrait passer par un étiquetage en parties du discours et une lemmatisation automatiques. Cet enrichissement soulève-t-il de trop nombreux problèmes ?

Ainsi, nous mettons en lumière un paradoxe : d'un côté, le travail est grandement facilité par la technique, d'un autre, aucune étude récente n'a vu le jour. Il y a bien évidemment des aspects méthodologiques et théoriques qui doivent être pris en compte, touchant notamment à la forme que doivent avoir les corpus. Mais on peut tout de même se demander si l'attente d'un corpus correspondant à un très haut degré d'exigence ne freine pas le travail sur corpus tout court et, plus problématique encore, s'il n'est pas indispensable de commencer par travailler sur des données regroupées de manière quelque peu opportuniste pour être en mesure de faire une typologie des

genres. D'autant que, lorsque l'on voit les matériaux à partir desquels la plupart des études précitées ont élaboré leurs descriptions, les données actuellement disponibles sont tout de même plus adaptées. Dans cette optique, nous avons pris la décision de constituer une archive provisoire à partir des données existantes.

3. CHOIX DES DONNÉES

3.1. L'existant

Il existe de nombreux travaux "théoriques" portant sur l'élaboration des corpus. Ils ont donné lieu à des recommandations[4] extrêmement précieuses concernant :

- Le recueil et les pièges de la transcription
- Le lien entre données et étude envisagée
- La documentation (métadonnées)
- Les principes de regroupement (typologie, genre, registre)
- Etc.

Mais ce que nous constatons, c'est que dans un grand nombre de travaux sur le français les chercheurs sont obligés de recourir à un compromis entre les ressources disponibles (en nombre limité) et l'étude qu'ils envisagent. La collecte des données respectant ces recommandations, bien que facilitée par la technologie, n'en reste pas moins une tâche ardue, surtout si l'on a besoin d'une quantité importante, comme c'est le cas lorsque l'on adopte une approche grammaticale ou lexicale. De plus, il n'y a aucune garantie que les regroupements effectués a priori soient les plus pertinents. Ce sont sans doute ces paramètres qui expliquent en partie pourquoi il existe si peu de travaux sur le français parlé dans une perspective grammaticale.

Compte tenu du travail colossal que représente la constitution d'un corpus oral, il paraît difficile d'élaborer à la fois l'intégralité des données et la grammaire, et ce dans des délais raisonnables. D'autant que nous retenons comme critère pertinent le principe qu'il faudrait disposer minimalement de deux tranches (une orale et une écrite) afin de pouvoir comparer et éclairer les résultats. En effet, nous considérons qu'une étude sur corpus n'a de sens que si elle s'insère dans une approche contrastive. Cependant, nous aurions préféré qu'elle soit doublement contrastive, entre l'écrit et l'oral et à l'intérieur de chaque canal, pour ne pas faire comme si écrit et oral représentaient deux entités totalement homogènes et opposées.

Pour constituer la tranche d'écrit et celle d'oral, nous nous sommes donc orienté vers la récupération de données existantes, comme l'ont fait les promoteurs du *Français Fondamental*. Malheureusement, plusieurs obstacles

4 Voir à ce sujet *Le guide des bonnes pratiques* et *Developing Linguistic Corpora: a Guide to Good Practice* à l'adresse suivante : http://ahds.ac.uk/creating/guides/linguistic-corpora/.

se sont à nouveau dressés sur notre chemin. Les ressources "disponibles" sont :

- Majoritairement non libres ;
- Déséquilibrées : beaucoup de littérature (dans la base de données *Frantext* notamment) et beaucoup moins de français parlé non planifié ;
- Éparpillées : il n'existe pas de "corpus de référence" ou d'archives dans lesquels nous irions piocher, comme on peut le faire avec le *British National Corpus*.

Heureusement, le paysage des corpus francophones est actuellement en profonde mutation et l'on assiste à des évolutions rapides. Les projets de mutualisation des ressources se multiplient. Pour le français parlé, on peut citer notamment les initiatives suivantes, qui mettent tout ou partie des données en téléchargement libre :

- Phonologie du Français Contemporain
- Centre de Ressources des Données Orales
- CLAPI
- Corpus de Français Parlé Parisien

D'autres devraient suivre dans les mois à venir :

- Centre National de Ressources Textuelles et Lexicales
- Enquête Sociolinguistique à Orléans
- Corpus de Référence du Français Parlé

Du coup, si l'on fait le cumul de toutes les ressources orales accessibles aujourd'hui, cela représente tout de même plusieurs millions de mots, ce qui permet d'envisager l'élaboration de notre grammaire, du moins au niveau quantitatif. Et ce nombre devrait rapidement progresser.

Mais la taille ne fait évidemment pas tout. Disposer d'une masse de données conséquente est nécessaire mais pas suffisant. Tout regroupement ne fait pas sens et on n'est pas à l'abri de récupérer des transcriptions non compatibles entre elles ou comportant de trop nombreuses erreurs.

3.2. Le compromis inévitable

En effet, quel corpus engendrerait le rassemblement de tout ce que nous pouvons récupérer (cf. Cappeau & Gadet, 2007) ? Le "mélange des genres" qui en résulterait pourrait invalider de fait les résultats obtenus, car les études sur le français parlé ont montré une très forte disparité des phénomènes linguistiques en fonction des types de données (Bilger & Cappeau, 2004 ; Cappeau, 2001).

En conséquence, nous rejetons la procédure consistant à récupérer, sans réflexion préalable et vérification minimale, tout ce qui est disponible et nous nous orientons vers une sélection minimale des données. En effet, les métadonnées sont parfois insuffisantes (ce qui peut rendre la transcription

inutilisable), les conventions de transcription divergentes et les formats de données hétérogènes. Un important travail de normalisation des données sera donc nécessaire pour pouvoir les mutualiser.

De plus, les principes minimaux que nous avons choisis de suivre permettent de remonter à chaque transcription. Si une transcription donnée présente un fonctionnement singulier, nous avons la possibilité de le savoir et de vérifier sur des données proches (si nous en disposons) les raisons de cette singularité (locuteur, thème abordé, etc.). Chaque transcription est donc conçue comme une entité indépendante.

Pour appuyer la pertinence de travailler à partir de données hétérogènes, on peut citer les nombreux travaux des membres du Groupe Aixois de Recherche en Syntaxe et de leurs continuateurs qui ont démontré l'intérêt que l'on pouvait tirer d'études portant à la fois sur des regroupements "tout venant" et des données plus homogènes (cf. Blasco-Dulbecco & Cappeau, 2004). Et le *British National Corpus*, bien qu'étant très hétérogènes dans sa partie "démographique", représente encore de nos jours un outil inestimable pour tous ceux qui souhaitent étudier l'anglais. Rappelons également que pendant de nombreuses années les travaux sur corpus pour le français ont porté essentiellement sur le journal *Le Monde* ou sur des textes littéraires et que les modèles de langage, même pour l'oral, ont été élaborés à partir de ces mêmes données. La prise en compte d'oral authentique pour la description grammaticale, appuyée par des méthodes modernes, représente déjà une avancée importante.

Qu'il s'agisse de l'écrit ou de l'oral, nous avons donc fait le choix de la diversité, en mettant tout de même des limites[5]. La récupération des données rend l'identification de chaque transcription fondamentale, afin de retrouver aisément leurs concepteurs et de visualiser les parties dans lesquelles un phénomène particulier se manifesterait. Cela permettra également d'écarter, le cas échéant, des transcriptions trop singulières et d'effectuer des regroupements ultérieurs.

Par exemple, dans un corpus de discours d'hommes et de femmes politiques sur lequel nous avons travaillé, nous avons remarqué un pic d'emploi de la préposition *contre* dans les discours d'A. Laguiller[6]. A elle seule, elle comptabilise plus de *contre* que les trois autres personnalités politiques réunies, bien que les tranches soient équilibrées en taille. De même, l'emploi de *lutte* comme recteur des syntagmes en *contre* est majoritaire, sauf dans les discours d'A. Laguiller, qui emploie plus souvent *attaque*. Et dans une étude comparative sur l'emploi de *lorsque* et *quand* au cours des siècles (cf. Benzitoun, 2006), basée sur *Frantext*, nous avions montré qu'une fréquence singulière, dans un siècle entier, n'était liée qu'à un seul auteur. Ne pas se restreindre aux fréquences globales et conserver la trace des partitionnements sont donc des paramètres indispensables pour toute étude sur corpus.

---

5 Pour une présentation des données retenues, se reporter à la section 4.

6 À proportion égale, A. Laguiller emploie 7 fois plus de *contre* que F. Mitterrand.

3.3. Données enrichies ou brutes ?

La question des informations à ajouter aux données primaires est également primordiale. Elle est d'autant plus importante que les étiquettes supplémentaires pourraient affecter nos résultats si elles se révélaient majoritairement fausses ou inadéquates. Cependant, le gain de temps lié à l'ajout automatique des étiquettes n'est pas négligeable. Et il est bien évident que sur les masses de données existantes cela permet de mener des études inenvisageables manuellement. Mais cela ne doit pas se faire au prix d'une trop grande détérioration de la qualité des résultats, comme nous allons le voir maintenant.

Dans la plupart des cas, ce que l'on appelle un corpus enrichi est un regroupement de textes segmentés en mots auxquels sont associés systématiquement la partie du discours (accompagnée éventuellement d'informations morphologiques) et le lemme correspondants. Cela peut également inclure des informations de nature sémantique ou syntaxique (dépendances, constituants, etc.), mais la technologie ne nous semble pas assez mature pour être utilisée, surtout lorsque l'on travaille sur des transcriptions d'oral non planifié. Afin d'effectuer nos tests, nous avons préféré nous limiter à des textes écrits pour ne pas rendre la tâche trop complexe, les étiqueteurs morphosyntaxiques ayant été initialement élaborés pour analyser de l'écrit.

Nous avons fait nos essais avec le logiciel libre *TreeTagger*, qui fournit les résultats sous la forme suivante :

```
je          PRO:PER          je
crois       VER:pres         croire
que         KON              que
vous        PRO:PER          vous
l'          PRO:PER          la/le
avez        VER:pres         avoir
traité      VER:pper         traiter
ça          PRO:DEM          cela
dans        PRP              dans
une         DET:ART          un
chronique   ADJ              chronique
```

Figure 1. Résultats fournis par *TreeTagger*

ainsi que *Cordial Analyseur* (logiciel payant sous licence) :

```
je          je          PPER1S
crois       croire      VINDP1S
que         que         SUB
vous        vous        PPER2P
l'          le          PPER3S
avez        avoir       VINDP2P
traité      traiter     VPARPMS
ça          ça          PDS
dans        dans        PREP
une         un          DETIFS
chroniqueL  chronique   NCFS
```

Figure 2. Résultats fournis par *Cordial Analyseur*

Le taux de précision de ces logiciels dépasse les 90 % mais, à notre connaissance, il n'existe pas d'étude poussée évaluant leur adéquation avec un véritable travail de description linguistique. C'est ce que nous nous proposons de faire dans la suite de cet article.

Afin de mener une première expérimentation, nous avons choisi des discours de François Mitterrand (environ 100.000 mots) et la forme *certain(e)s* uniquement au pluriel. Le classement obtenu est le suivant :

Etude manuelle	Treetagger	Cordial
DET 51	ADJ 3	ADJ 56
PRO 23	PRO:IND 71	PRO 18

Si *Cordial Analyseur* donne des résultats assez proches de ce que l'on obtient manuellement, les étiquettes apposées par *TreeTagger* sont en très grande partie fausses. Ainsi, il est patent que l'utilisation de *TreeTagger* est à proscrire pour ce type d'études.

Cependant, nous avons souhaité aller plus loin dans l'évaluation de *Cordial*. Dans la suite, nous prenons également en compte le genre et le nombre de *certain* et nous passons à un corpus plus vaste composé exclusivement de presse écrite (environ 270.000 mots).

Les étiquettes apposées sont les suivantes avec les exemples correspondants :

Adjectif féminin singulier (ADJFS)	*Une certaine humeur*
Adjectif masculin singulier (ADJMS)	*Un certain équilibre*
Adjectif masculin pluriel (ADJMP)	*Des dividendes certains*
Adjectif indéfini (ADJIND)	*Certaines entreprises*
Pronom indéfini féminin pluriel (PIFP)	*Certaines sont tombées*
Pronom indéfini masculin pluriel (PIMP)	*Certains ont refusé*

Voici les résultats obtenus :

	MANUEL	CORDIAL
ADJFS	35	31
ADJIND	113	**140**
ADJMP	1	1
ADJMS	30	28
PIFP	8	**2**
PIFS	0	1
PIMP	53	**37**

Le taux d'erreurs est nettement plus visible dans ce contexte. Sur 240 occurrences, 30 sont mal étiquetées (soit 1 exemple sur 8) et 4 catégories sont sous-évaluées. Sur ces 30 erreurs, 28 sont liées à une classification erronée comme adjectifs indéfinis. L'erreur de loin la plus courante (24 cas sur 30) est le transfert des pronoms indéfinis dans la catégorie des adjectifs indéfinis. Le calcul de la fréquence des adjectifs génèrera donc le plus de bruit (et, en conséquence, une surreprésentation de cette catégorie) et celui des pronoms le plus de silence. Sur 61 occurrences possibles de *certain(e)s* pronom, seulement 38 ont été correctement retrouvées (40 dont deux erreurs). Et si l'on souhaite par exemple étudier les emplois de *certain(e)s* directement en position de sujet ou de complément, il va sans dire que les résultats seront problématiques. Et la même étude sur des corpus oraux risque de faire penser à tort qu'il n'existe pas de *certain* en position sujet ou complément, étant donné leur faible proportion dans ce type de données (cf. Cappeau).

Ainsi, nous rejoignons en partie la position de J. Sinclair pour qui l'étiquetage transforme en profondeur les données (on ne travaille plus sur le texte mais sur les étiquettes) et peut être vu plus comme une contamination que comme un enrichissement (cf. Sinclair, 2004).

4. CONCLUSION

À la fin de ce rapide parcours, nous sommes désormais en mesure de présenter l'archive à partir de laquelle nous allons travailler. Nous préférons parler d'archive plutôt que de corpus, car nous n'avons pas scrupuleusement suivi les principes fondamentaux que doit respecter une collection de données pour prétendre à l'appellation de corpus. Nous avons tout de même décidé d'instaurer un critère de regroupement relativement peu contraignant, mais fort utile, à savoir le degré de planification. La partie orale est composée de productions majoritairement non planifiées et la partie écrite de productions élaborées. C'est le trait [+/- planifié] qui est pertinent et non une opposition trop simpliste entre oral et écrit. En outre, nous souhaitions que l'archive soit d'une taille importante, tout en restant accessible à une exploitation manuelle des concordances pour pouvoir contrôler chaque étape.

Pour l'oral, elle est composée du *Corpus de Français Parlé Parisien* (CFPP2000), de *Corpaix* dans sa version de mai 2000, du *Corpus de Référence du Français Parlé*, et d'une partie de la base *Phonologie du Français Contemporain*. Plusieurs paramètres ont été pris en compte dans ce choix : données à notre disposition, libres pour certaines, le plus possible composées de productions non planifiées, productions d'adultes ou d'adolescents et dans une aire géographique limitée, à savoir la France. Ainsi, dans PFC, nous n'avons récupéré que les discussions libres et seulement de locuteurs français. Toutefois, nous avons retenu, de manière minoritaire, des transcriptions ne respectant pas certains de ces critères, afin de voir si ces données se singularisent dans le cadre des descriptions que nous ferons. La totalité des données orales fait un peu moins de deux millions de mots (1.950.000 pour être précis).

Pour l'écrit, nous avons également opté pour un empan assez large, tout en ne retenant que des textes ayant demandé un certain degré de planification. Pour le constituer, nous avons sélectionné des tranches dans le *Corpus Evolutif de Référence du Français*, à savoir :

- De la presse : Courrier international, Le Monde, Le Monde diplomatique, le nouvel Observateur ;
- Des discours politiques : J. Chirac, L. Jospin, F. Mitterrand ;
- Des textes scientifiques : CNRS éditions, Revues Hermès, Pour la science, Sciences et avenir ;
- Des textes institutionnels : Assemblée Nationale, textes juridiques et législatifs ;
- Divers autres textes : Philosophie, critiques littéraires, critiques cinéma, nouvelles de science fiction, romans.

Ce recueil totalise 1.945.000 mots, ce qui est comparable à la partie orale. C'était une autre de nos exigences à laquelle nous avons pu répondre.

Ces regroupements sont en partie opportunistes et hétérogènes. Mais notre archive n'est pas figée une fois pour toute. Elle fera l'objet d'ajustements si nécessaire, lorsque nous lancerons nos descriptions. Cela permettra d'améliorer nos connaissances concernant la notion de genre, car c'est en élaborant des descriptions sur des données "tout venant" que nous pourrons effectuer d'éventuels regroupements. C'est de cette manière que nous ferons avancer la recherche sur la typologie des données. De toute façon, compte tenu des connaissances actuelles, il semble difficile d'envisager des regroupements opératoires en amont. De même, la limitation en taille pourra être compensée par le recours à des données externes pour vérifier la distribution de fréquences remarquables.

En outre, nous avons décidé de ne pas avoir recours à une annotation automatique en parties du discours pour les raisons signalées plus haut. Nous aurions pu procéder à une correction systématique, mais il s'agit d'un travail long pour lequel nous ne sommes pas sûr de récolter tous les bénéfices escomptés. De plus, les parties du discours étant théoriquement marquées, cela

aurait eu comme inconvénient de rendre difficile la perspective de "corpus driven" dans laquelle nous travaillons. Il est également possible que travailler à partir de corpus annotés nous fasse passer à côté d'exemples presque invisibles à l'œil nu, mais qui pourtant revêtent une très grande importance. Cependant, nous n'avons pas encore procédé à des tests concernant la lemmatisation. Il est possible que nous retenions ce type d'information supplémentaire. Pour l'instant, notre corpus est donc "brut".

Les données à partir desquelles nous allons mener nos descriptions, bien qu'imparfaites, représentent tout de même une avancée significative par rapport aux exemples relevés à la volée ou extraits de transcriptions en nombre très limité qui existaient auparavant. Mais seul un grand nombre de travaux permettra de valider notre démarche, éventuellement de procéder à des regroupements et de "nettoyer" les données. Nous envisageons donc la phase d'exploitation de pair avec celle de constitution.

Nous disposons désormais des données minimales pour jeter les bases de notre grammaire sur corpus. Elles sont pour l'instant dans un état transitoire, en attendant d'évaluer le caractère opératoire de notre approche. La phase de normalisation, tant des données que des métadonnées, n'aura donc lieu qu'à la condition que nos choix soient validés par des études descriptives. Et après accord des concepteurs des corpus oraux, nous envisageons de mettre à disposition les parties libres de droit sous une forme comparable. Ainsi, cela permettra de disposer de données à partir desquelles une communauté entière pourra travailler, ce qui ne peut être que bénéfique pour la structuration de la linguistique française sur corpus.

BIBLIOGRAPHIE

BAUCHE H. (1928), *Le français populaire*, Paris, Payot.

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

BENZITOUN C. (2006), *Quand* et *lorsque* sont-ils synonymes ?, *Working Papers in Linguistic Informatics*, 12, 167-182.

BIBER D., JOHANSSON S., LEECH G., CONRAD S. & FINEGAN E. (1999), *Longman grammar of spoken and written English*, London, Longman.

BILGER M. & CAPPEAU P. (2004), "L'oral ou la multiplication des styles", *Langage & société,* 2004/3, 109, 13-30.

BLANCHE-BENVENISTE Cl., BILGER M., ROUGET C., EYNDE K. van den & MERTENS P. (1990), *Le français parlé : études grammaticales*, collection Sciences du langage, éditions du CNRS.

BLANCHE-BENVENISTE Cl. (2000), "Corpus de français parlé", in Bilger M. (éd.), *Corpus - Méthodologie et applications linguistiques*, Paris. Honoré Champion.

BLANCHE-BENVENISTE Cl. (2000), "Le français parlé : un regard sur la syntaxe", dans Antoine G. & Cerquiglini B. (éds), *Histoire de la langue française 1945-2000,* Paris, CNRS Editions, 195-197.

BLASCO-DULBECCO M. & CAPPEAU P. (2004), "Quelques remarques sur l'adjectif à l'oral", *L'adjectif en français et à travers les langues*, Caen, Presses universitaires de Caen, 1-16.

CAPPEAU P. & GADET F. (2007), "L'exploitation sociolinguistique des grands corpus. Maître-mot et pierre philosophale", *Revue Française de Linguistique Appliquée*, 2007/1, 121, 99-110.

CAPPEAU P. (2001), "Faits de syntaxe et genres à l'oral", *Le français dans le monde*, numéro spécial, 69-77.

DAMOURETTE J. & PICHON E. (1911-1946), *Des mots à la pensée. Essai de grammaire de la langue française*, Vrin.

DELIC (2004), Présentation du *Corpus de Référence du Français Parlé*, *Recherches sur le français parlé*, 18, 11-42.

DURAND J., LAKS B. & LYCHE C. (2002), "La phonologie du français contemporain : usages, variétés et structure", in Pusch C. & Raible W. (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", in Williams G. (éd.), *La linguistique de corpus*, Rennes, Presses Universitaires de Rennes, 205-217.

FRANÇOIS D. (1974), *Français parlé : analyse des unités phoniques et significatives d'un corpus recueilli dans la région parisienne*, 2 volumes, Société d'études linguistiques et anthropologiques de France, Paris.

GADET F. (2007), *La variation sociale en français*, Paris, Ophrys.

GOUGENHEIM G., MICHEA R., RIVENC P. & SAUVAGEOT A. (1964), *L'élaboration du français fondamental : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris, Didier.

HABERT B. (2000), "Des corpus représentatifs : de quoi, pour quoi, comment", Bilger M. (éd.), *Linguistique sur corpus - Etudes et réflexions*, Presses Universitaires de Perpignan, 11-58.

MARTINON P. (1927), *Comment on parle en français : la langue parlée correcte comparée avec la langue littéraire et la langue familière*, Paris, Larousse.

MILLON C. & ANTONIOTTI M. (2002), "Une expérience de constitution d'un corpus de référence du français contemporain sur le Web", Colloque *Corpus et Web*, 26-27 novembre, Saint-Denis.

RASTIER F. (2005), "Enjeux épistémologiques de la linguistique de corpus", in Williams G. (éd.), *La linguistique de corpus*, Presses universitaires de Rennes, 31-45.

RIGAULT A. (1971) (dir.), *La grammaire du français parlé*, Recherches/Applications, Hachette.

SANDFELD K. (1936), *Syntaxe du français contemporain,* 2 tomes, Copenhague-Paris, Librairie E. Droz.

SINCLAIR J. (1997), *Collins Cobuild English Dictionary*, London, Harper Collins.

SINCLAIR J. (2004), *Trust The Text: Language, Corpus and Discourse*, Routledge.

Les corpus sources

Corpus de Français Parlé Parisien (CFPP)
S. Branca-Rosoff, S. Fleury, F. Lefeuvre, M. Pires
Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000)
Adresse pour le consulter : http://ed268.univ-paris3.fr/CFPP2000/

Phonologie du Français Contemporain (PFC)
Durand, Jacques, Bernard Laks & Chantal Lyche
Adresse pour le consulter : http://www.projet-pfc.net

Corpus de Référence du Français Parlé (CRFP)
Ce corpus répond à une requête de la Délégation à la langue française (Ministère de la Culture), qui l'a totalement financé. La réalisation de ce projet avait été confiée, en 1998, à l'équipe *Corpus* de l'Université de Provence, dirigée par Claire Blanche-Benveniste et associée au CNRS. À partir de 2000, le projet a été pris en charge par l'équipe DELIC (*DEscription Linguistique Informatisée sur Corpus*), dirigée par Jean Véronis.
Adresse pour le consulter : aucune car non distribué.

Corpus Evolutif de Référence du Français (CERF)
Corpus élaboré sous la direction de J. Véronis à l'Université de Provence.
Adresse pour le consulter : aucune car non distribué.