Présentation du *Corpus de référence du français parlé*

Équipe DELIC

Résumé

Le *Corpus de référence du français parlé*, qui vient d'être réalisé par l'équipe DELIC, vise à mettre à la disposition de la communauté des linguistes, chercheurs et enseignants, un témoignage de la langue française parlée aujourd'hui dans les principales villes de l'hexagone. Le corpus comporte 440 000 mots, correspondant à plus de 36 heures de parole. Il est composé de 134 enregistrements échantillonnés en fonction de plusieurs situations de parole et de niveaux d'études des locuteurs. La totalité du corpus se présente sous une forme transcrite alignée avec le son.

1. Introduction

Le *Corpus de référence du français parlé* que nous présentons ici (et qui a été utilisé dans la plupart des articles de ce numéro) 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. A partir de 2000, le projet a été pris en charge par l'équipe DELIC (DEscription Linguistique Informatisée sur Corpus), dirigée par Jean Véronis.

L'objectif de ce corpus est de mettre à la disposition de la communauté des linguistes, chercheurs et enseignants, un témoignage de la langue française parlée aujourd'hui dans les principales villes de l'hexagone. Il s'agissait avant tout de recueillir des données représentatives d'un français parlé que nous pourrions qualifier d'« usage général et

Recherches sur le français parlé, 18, 11-42

courant », ce qui nous a amenés à effectuer certains choix touchant aussi bien aux caractéristiques des locuteurs qu'aux situations de parole.

Ce corpus compte environ 440 000 mots et il est constitué de 134 enregistrements dont la partie transcrite correspond à une durée moyenne de 16 min. 48 s[1]. La totalité représente 36 heures et 50 minutes de parole. Dans le projet initial, il était prévu que le corpus comporte 160 enregistrements, mais les problèmes de coordination et d'harmonisation, dus à la multiplicité des régions géographiques, des enquêteurs et des transcripteurs, avaient été sous-estimés. Nous n'avons pas voulu sacrifier la qualité du résultat, ni accroître exagérément les délais de réalisation, car ce projet représente une occasion unique de fournir à la communauté de recherche francophone un outil scientifique important, qui lui fait cruellement défaut, alors que des corpus oraux importants existent ou sont en cours de réalisation pour la plupart des langues européennes. Pour le français parlé, il offre une base de comparaison avec les corpus de français parlé hors hexagone, le corpus Valibel en Belgique et le corpus d'Ottawa-Hull au Canada.

2. Échantillonnage

Nous décrivons ici les principes d'échantillonnage qui ont présidé à la construction du corpus, et nous fournissons un certain nombre de données quantitatives qui permettent de se faire une idée de la diversité des données disponibles[2].

2.1. Répartition géographique

Les données ont été recueillies dans 37 villes de province (27 dans la zone nord et 20 dans la zone sud), et en région parisienne. Les villes de province sont des villes de dimension moyenne comme Perpignan ou Pau, ou beaucoup plus grandes comme Bordeaux ou Lyon. Seules deux villes ont moins de 10 000 habitants (Corte et Saint-Affrique) (Figure 1). La

[1] Les enregistrements étaient généralement plus longs mais seule une partie a été transcrite. Cette partie n'est pas nécessairement le début de l'enregistrement, de façon à éviter des fragments liés à la mise en place de l'interaction (« bon, alors, ça enregistre ? », « alors, de quoi je parle ? », etc.).

[2] Les critères d'échantillonnage adoptés pour la constitution de ce corpus ont été élaborés par Mireille Bilger.

région parisienne a été divisée en 5 zones : Paris-centre et les secteurs nord-ouest, nord-est, sud-ouest et sud-est.

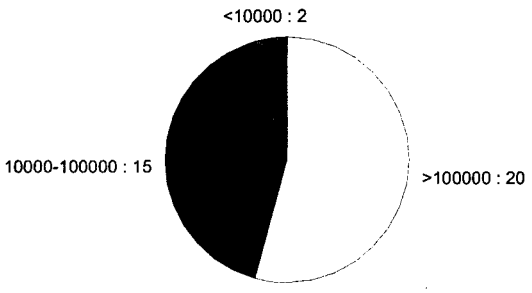


Figure 1. Répartition des villes de province en fonction de la population

Au total, 107 enregistrements ont été effectués en province (50 dans la zone nord, 57 dans la zone sud) et 27 enregistrements dans la région de Paris et sa banlieue (Figure 2). Nous obtenons une moyenne de 3 enregistrements par ville, sauf pour la région parisienne pour laquelle nous avions envisagé dès le début un plus grand nombre d'enquêtes ; chaque secteur compte entre 5 et 6 enregistrements (Figure 3).

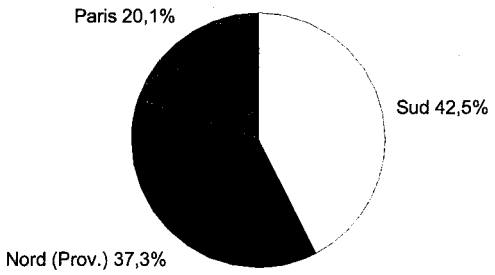


Figure 2. Répartition géographique des enregistrements

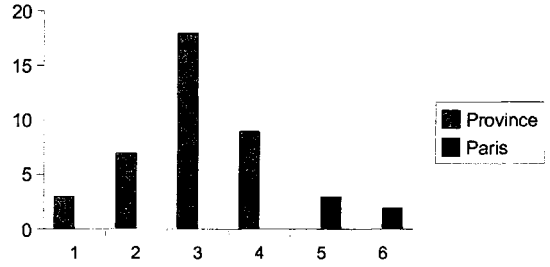


Figure 3. Nombre d'enregistrements par ville

2.2. Taille des transcriptions

Comme il a été dit en introduction, la durée moyenne des transcriptions est de 16 min. 48 s[3]. La plupart des transcriptions se situent autour de cette valeur, mais quelques transcriptions sont notablement plus courtes ou plus longues (de 6 min. 45 s. à 31 min. 16 s.). La Figure 4 donne la distribution des durées des transcriptions.

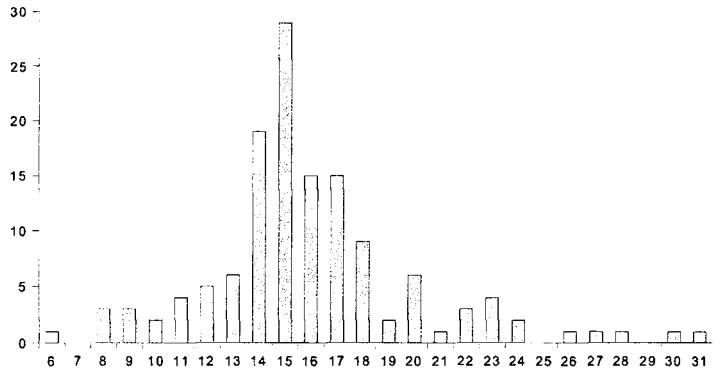


Figure 4. Distribution des durées des transcriptions

[3] Les valeurs que nous donnons tiennent compte des interruptions du discours (musique, déplacement du locuteur, etc.), dont la durée a été déduite de la durée totale de l'enregistrement.

2.3. Répartition selon la situation de parole

Au critère géographique s'ajoute la prise en compte de la situation de parole. Nous avons défini trois situations d'enregistrement :

- **la parole privée** : sous la forme d'un entretien sollicité spécifiquement dans le cadre de l'enquête, cette situation de parole renvoie à deux types de productions : le récit de vie (dont le contenu peut varier : récit d'un voyage, d'une expérience, souvenirs d'enfance, etc.), ou la présentation d'un « savoir-faire » professionnel ou autre.
- **la parole professionnelle** : entretiens également sollicités spécifiquement, mais dans lesquels les locuteurs ont été enregistrés dans l'exercice de leur fonction ou quand ils parlent de leur profession sur leur lieu de travail.
- **la parole publique** : cette situation se distingue des deux autres par le fait que les intervenants s'expriment toujours en présence d'un public ; elle comporte une partie d'entretiens sollicités, le reste étant constitué d'émissions radiophoniques (actualités, interview, table ronde, tribune téléphonique, etc.), de cours et conférences, de réunions politiques ou associatives (conseil municipal, discussion syndicale, comité de quartier, etc.), et de quelques situations plus spécifiques (visite de musée, dégustation de vins, etc.).

Remarquons que la distinction entre les types « privé » et « professionnel » n'est pas subordonnée à une différence de contenu, puisque le fait pour un locuteur de parler d'un savoir-faire professionnel est représenté dans l'un comme dans l'autre type : simplement, on a tenu à considérer de manière indépendante les échanges qui se situent sur le lieu professionnel dans la mesure où la parole des locuteurs peut y prendre une forme plus « institutionnelle ».

La répartition sur l'ensemble du corpus est la suivante :

Type	**Enregistrements**		**Mots**	
Privé	84	62,7%	282857	64,5%
Prof.	22	16,4%	75001	17,1%
Public	28	20,9%	80601	18,4%
Total	**134**	**100,0%**	**438378**	**100,0%**

On voit que la parole « privée » représente environ les deux tiers du corpus, tandis que les types « professionnel » et « public » se partagent le tiers restant à parts à peu près égales (Figure 5).

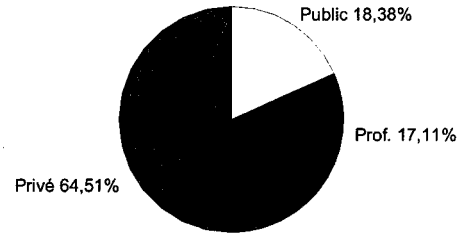


Figure 5. Répartition par type de parole (en nombre de mots)

La parole « publique » se décompose de la façon résumée par la Figure 6. On notera que les enregistrements de parole « publique » comportent trois entretiens-enquêtes réalisés en public. Ces entretiens s'ajoutent à ceux des sous-corpus de parole privée (84 entretiens) et professionnelle (22 entretiens) ; au total, le corpus comprend donc 84 + 22 + 3 = 109 entretiens, ce qui représente 81% des enregistrements.

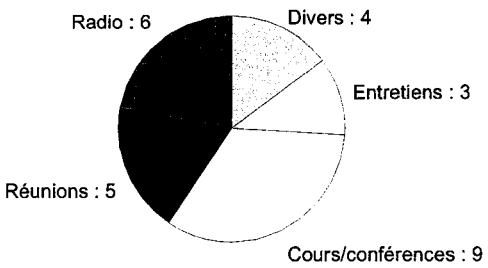


Figure 6. Situations de parole publique

2.4. Nombre de locuteurs

Le nombre de locuteurs[4] s'échelonne entre 1 et 18, selon la répartition donnée par la Figure 7. Le nombre moyen de locuteurs est de 2,7 ; la situation la plus fréquente fait intervenir 2 locuteurs.

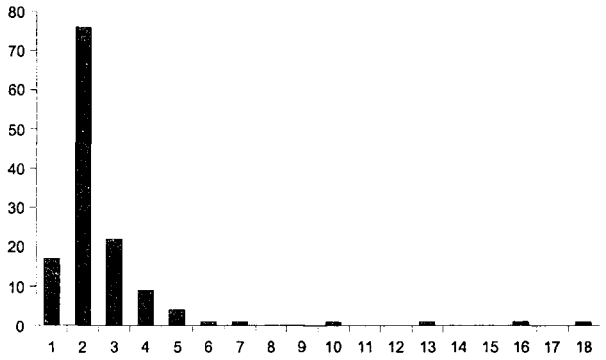


Figure 7. Répartition par nombre de locuteurs

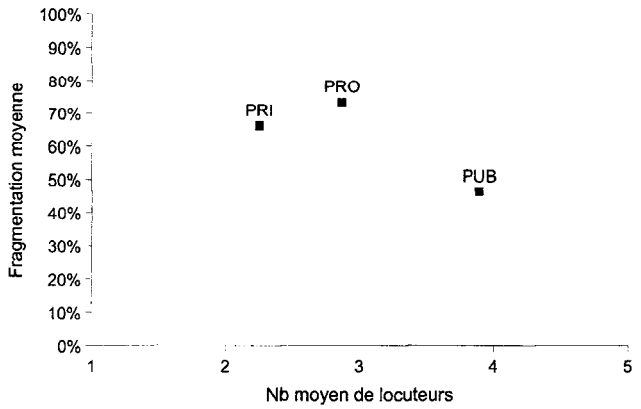
Si ce paramètre permet d'apprécier la part respective de monologues (17%) ou de situations faisant intervenir des locuteurs nombreux (19% des transcriptions font intervenir 4 locuteurs et plus), il reflète mal la répartition réelle de la parole. Par exemple, une situation à deux locuteurs peut être une situation d'entretien dans laquelle l'un des locuteurs a un usage prépondérant de la parole, ou une situation de discussion, où les prises de parole sont plus équilibrées. De même, une situation peut faire intervenir trois locuteurs, mais l'un d'entre eux peut avoir un rôle extrêmement marginal (par exemple traverser la pièce et dire bonjour).

Nous avons donc défini un autre paramètre, de fragmentation de la parole, qui permet de mieux quantifier la répartition de la parole entre locuteurs. Le taux de fragmentation varie entre 0 (parole totalement monopolisée par un locuteur ou monologue) et 100% (parole fragmentée de

[4] Dans la partie transcrite. Il peut arriver que d'autres locuteurs interviennent dans une autre partie de l'enregistrement. Ils n'ont pas été comptabilisés.

façon maximale entre locuteurs, dans le cas théorique de prises de parole en nombre égal et de même longueur)[5].

La Figure 8 représente le nombre moyen de locuteurs et le taux de fragmentation moyen dans chacun des sous-corpus. On notera le taux important de fragmentation des entretiens : on aurait pu attendre des situations asymétriques, où la parole revient de façon prépondérante au locuteur interviewé, mais on s'aperçoit que le taux de fragmentation est important, la plupart des entretiens se situant dans la zone de 75% à 100%, qui indique que l'enquêteur intervient largement dans la discussion[6]. En ce qui concerne le sous-corpus « public », un examen plus approfondi des données révèle une distribution bi-modale, à mettre en relation avec son caractère composite, comportant à la fois des situations de parole très fragmentées (réunions) et des situations proches du monologue (par exemple cours).



5 Ce paramètre se calcule de la façon suivante :

$$f = 1 - \frac{\sigma/m}{\sqrt{N-1}}$$

N étant le nombre de tours de parole, m le nombre moyen de mots par tour de parole et σ l'écart-type correspondant.

6 Le taux de fragmentation ne mesure que la part relative de la parole, d'un point de vue strictement quantitatif, sans préjuger du contenu de l'interaction. On notera par exemple, de ce point de vue, que la transcription la plus fragmentée (PRI-SAI-1) est un entretien, dans lequel les interventions de l'enquêteur sont presque exclusivement composées de « oui oui oui », « ah d'accord », etc.

Figure 8. Fragmentation de la parole

2.5. Age, sexe et niveau scolaire des locuteurs (entretiens)

Ces différents paramètres n'ont pu être contrôlés que pour les entretiens sollicités dans le cadre de l'enquête. Les indications suivantes portent donc sur les 109 enregistrements de cette catégorie, en ne considérant que le locuteur principal (l'interviewé).

Les tranches d'âge se répartissent ainsi (voir aussi Figure 9) :

Age	Transcriptions
18-30 ans	33
30-65 ans	55
+ 65 ans	21
Total	**109**

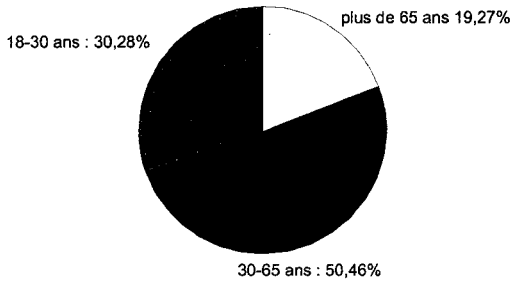


Figure 9. Répartition des entretiens par tranche d'âge

La répartition par sexe ne faisait pas partie des critères initiaux, mais elle a pu être contrôlée a posteriori. Les interviewés se répartissent en 44% de femmes et 56% d'hommes (Figure 10).

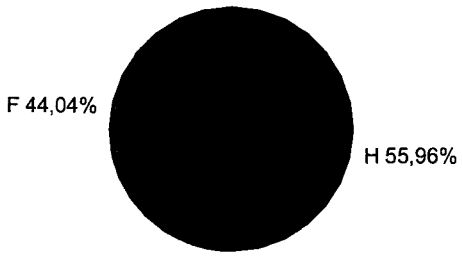


Figure 10. Répartition des entretiens par sexe

Trois tranches ont été définies en ce qui concerne le niveau scolaire des locuteurs : niveau primaire ou collège, niveau bac ou niveau enseignement supérieur (à partir de Bac+3). La répartition se présente ainsi (voir aussi Figure 11) :

Niveau	Transcriptions
Inconnu	6
Primaire-Collège	27
Bac	45
Supérieur	31
Total	**109**

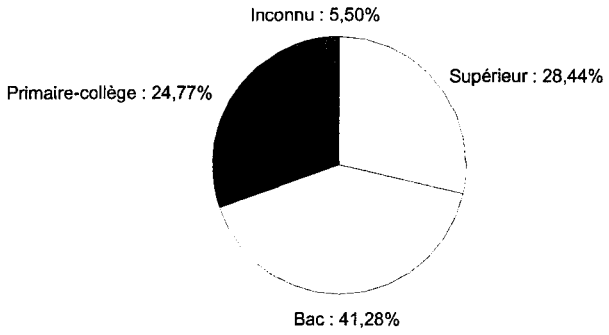


Figure 11. Répartition des entretiens par niveau scolaire

3. Recueil et transcription

3.1. Les enregistrements

Les enregistrements ont été effectués à l'aide de « baladeurs » mini-disques, sans faire appel à des microphones bidirectionnels ou à des micro-casques, ce qui a néanmoins permis une qualité sonore correcte dans la plupart des cas. Le but du corpus étant de recueillir du français non lu dans un cadre le moins contraint possible, la qualité sonore n'est pas forcément celle que peuvent exiger, par exemple, des études phonétiques précises. Les enregistrements ont été catégorisés en trois niveaux de qualité : A (excellent), B (bon) et C (passable). Le tableau ci-dessous montre la répartition dans le corpus :

Qualité	Enregistrements	
A	104	77,6%
B	16	11,9%
C	14	10,4%
Total	**134**	**100,0%**

Tableau 1. Répartition de la qualité des enregistrements du corpus

La numérisation, la délimitation des parties à transcrire ainsi que le réglage du volume ont été réalisés grâce au logiciel Sound Forge XP de la société Sonic Foundry[7].

3.2. La fiche signalétique

Chaque transcription est accompagnée d'une fiche signalétique qui donne les informations suivantes :

- identificateur de la transcription ;
- titre ;
- lieu d'enregistrement ;
- responsable de l'enregistrement ;
- transcripteur ;
- réviseur(s) ;
- type de parole ;
- situation ;

[7] http://www.sonicfoundry.com/

- résumé ;
- nombre de locuteurs ;
- numéro du locuteur principal (interviewé), le cas échéant ;
- nombre de mots ;
- durée ;
- particularités.

De plus pour chaque locuteur, la fiche donne un certain nombre de renseignements :

- numéro du locuteur
- âge
- sexe
- niveau d'études
- profession
- rôle (enquêteur, interviewé, etc.)

L'Annexe 1 donne un exemple de fiche signalétique.

3.3. La transcription

Les conventions utilisées pour la transcription du *Corpus de référence du français parlé* sont largement inspirées de celles mises au point par le G.A.R.S. (cf. Blanche-Benveniste et al., 1991: 228-229). Ces conventions préconisent une représentation en orthographe standard, sans ponctuation, ni indications prosodiques, signes démarcatifs ou expressifs. Les pauses sont notées, ainsi que les amorces (mots inachevés). On tient également à conserver les traces de certaines difficultés de transcription liées à des problèmes d'interprétation. Le ou les transcripteurs peuvent percevoir des séquences sonores différentes ou proposer plusieurs versions orthographiques pour une seule et même séquence sonore. Un certain nombre d'aménagements ont été apportés aux conventions du GA.R.S., principalement pour des raisons techniques, en particulier pour éviter des choix liés à des logiciels de traitement de textes particuliers, et permettre une plus grande robustesse et une meilleure portabilité. Par exemple, les chevauchements, qui étaient notés par un soulignement :

L1 de me penser à penser à trouver un travail - et l'imprimerie
L2 vous aviez à peu près quel âge

sont désormais notés par des chevrons ouvrants et fermants :

L1 de me penser à penser à trouver un travail + < et l'imprimerie
L2 vous aviez à peu près > quel âge

Pour la même raison, les transcriptions phonétiques ponctuelles (mots mal prononcés, etc.) ne sont plus notées à l'aide d'une police particulière, mais font appel à l'alphabet SAMPA (cf. Annexe 3).

D'autre choix visent à une amélioration ergonomique du travail de transcription, en évitant des confusions possibles de la part du transcripteur. Ainsi, le tiret servait précédemment aussi bien à marquer les amorces que les pauses :

alors nous le euh - nous avons a- accepté de - de les prendre avec nous

Toutefois, ces conventions étaient « fragiles » dans la mesure où une espace mal placée pouvait totalement transformer le codage. Les pauses sont désormais notées par le signe + (plus) :

alors nous le euh + nous avons a- accepté de + de les prendre avec nous

L'annexe 2 donne la liste des transcriptions révisées.

La transcription a été réalisée à l'aide du logiciel *Transcriber* (Barras et al. 1996)[8], qui a également permis de réaliser un alignement de la totalité du corpus avec le son. La segmentation a été faite aux pauses les plus importantes (la durée moyenne des segments est de 3,1 s. pour 9,6 mots). L'exemple ci-dessous montre un exemple d'alignement (la durée est exprimée en secondes) :

L1	0.0	voilà nous nous étions interrompu au moment où nous avons rencontré les deux Anglaises dont une était euh +
	7.3	noire +
	8.6	et qui était d'ailleurs la seule parlant français alors nous le euh +
	12.3	nous avons a- accepté de +
	14.1	de les prendre avec nous nous nous sommes mis d'accord pour partager les frais d'essence euh +
	18.9	parce que c'est vrai que ça faisait hum +
	22.5	ça fait l'éch- l'essence était très chère +
	25.1	et euh +
	27.3	quand on en trouvait +
	29.4	et nous voilà partis donc euh à cinq +
	33.4	le Dogon les deux Anglaises mon épouse et moi euh et euh +

[8] Logiciel téléchargeable gratuitement sur le site : http://www.etca.fr/CTA/gip/Projets/Transcriber/

38.7 et six avec le chauffeur même +
41.1 nous nous n'oublions pas ce +
44.0 le chauffeur +

Cet alignement très fin permet d'écouter à volonté les passages souhaités lors de l'examen des concordances. Il constitue un nouveau mode d'exploration des corpus et permet une vérification immédiate de chaque passage de la transcription.

3.4. Problèmes juridiques et anonymisation

Le recueil et la diffusion de corpus oraux pose des problèmes délicats du point de vue juridique, et l'équipe a pris deux mesures permettant d'éviter des difficultés ultérieures[9] :

- chaque personne enregistrée a signé une autorisation d'enregistrement et de diffusion, à l'exception des enregistrements radiophoniques pour lesquels une négociation est en cours ;
- tous les noms de tierces personnes, de sociétés ou de lieux pouvant poser difficulté (propos péjoratifs, etc.) ont été anonymisés aussi bien dans les enregistrements (remplacés par un bip sonore) que dans les transcriptions (remplacés par un code du type *P* pour les patronymes, etc.).

4. Exploitation et diffusion des données

Le corpus aligné peut être exploité à l'aide du concordancier *Contextes,* développé par Jean Véronis[10], qui possède de nombreuses fonctionnalités, et permet d'écouter les segments correspondant à chaque ligne de concordance.

Trois modes de diffusion sont envisagés à l'heure actuelle[11] :

- une version libre sur le World Wide Web, permettant une extraction limitée de concordances[12], sans accès au son ;

---

9 L'équipe est reconnaissante à Sandra Teston pour ses recherches sur les questions juridiques.

10 http://www.up.univ-mrs.fr/veronis/logiciels/Contextes

11 Ces dispositions peuvent changer en fonction des problèmes juridiques ou techniques rencontrés.

- une version orientée vers les enseignants et chercheurs en linguistique, distribuée sur CD-ROM comportant l'intégralité du son compressé au format WMA (Windows Media Audio), et permettant l'extraction de concordances à l'aide du logiciel *Contextes* ;
- une version destinée à la recherche sur les technologies de la parole, avec son au format WAV, et comportant les transcriptions au format XML utilisé par le logiciel *Transcriber.*

5. Conclusion

Il va de soi que, malgré l'effort considérable qu'a représenté sa constitution, le *Corpus de référence du français parlé* demande à être étendu : bien des situations de parole sont encore absentes, trop d'usages de la langue ne sont pas représentés. Néanmoins, en l'état actuel, le corpus se présente déjà comme une ressource majeure pour l'étude et l'enseignement du français. Par sa taille (440 000 mots, 36 heures d'enregistrement), et son large échantillonnage, il constitue une base de données comportant une variété importante d'accents, d'expressions, de tournures, qui donne une image linguistique précise de la France urbaine contemporaine. La présentation, dans laquelle le texte et le son sont alignés, offre aux utilisateurs de nouvelles possibilités d'exploitation et d'exploration du corpus.

Références

Barras, C., Geoffrois, E., Wu, Z. & Liberman, M. (2000). *Speech Communication* (Special issue on Speech Annotation and Corpus Tools). *33*(1-2) : 5-22.

Blanche-Benveniste, C., Bilger, M., Rouget, C. & van den Eynde, K. (1991). *Le français parlé. Études grammaticales.* Paris : CNRS Éditions.

[12] http://www.up.univ-mrs.fr/delic/crfp

Annexe 1. Exemple de fiche signalétique

Identificateur	PRI-AIX-1
Titre	Afrique
Lieu d'enregistrement	Aix-en-Provence
Responsable de l'enregistrement	André Valli
Transcripteur	Christophe Rey
Réviseur	André Valli
Type de parole	Privée
Situation	Entretien
Résumé	Récit d'un voyage au Mali
Nombre de locuteurs	2
Locuteur principal	L1
Nombre de mots	2903
Durée	00:15:17
Particularités	Néant

Locuteurs

L1	Age	30-65
	Sexe	H
	Niveau d'études	Bac
	Profession	Retraité
	Rôle	Interviewé
L2	Age	30-65
	Sexe	H
	Niveau d'études	Supérieur
	Profession	Professeur des universités
	Rôle	Enquêteur

Annexe 2 – Conventions de transcription

1. Principes généraux

La transcription se fait en orthographe standard, sans ponctuation ni majuscule de début de phrase. Chaque tour de parole fait l'objet d'un paragraphe séparé, qui commence par l'identification du locuteur.

> Attention, ne pas insérer une fin de paragraphe (retour chariot) à chaque bout de ligne d'écran, mais seulement aux fins de tours de parole).

Exemple :

L1 donc vous faites quoi dans la vie
L2 mais je travaille ici à la fac + euh je fais le ménage + je fais un peu dans le secrétariat le mercredi après-midi + à l'imprimerie quoi +
L1 et vous faites quoi dans le secrétariat
L2 ben je fais les photocopies {rire} + et je trie euh et j'agrafe +

Un certain nombre de symboles particuliers peuvent ou doivent être utilisés au cours de la transcription. Ils sont résumés dans le tableau ci-dessous, et leur utilisation est détaillée dans la suite du texte.

{...}	éléments métalinguistiques (commentaires, etc.)
[...]	prononciations particulières
< ... >	chevauchements de locuteurs
+	pauses
///	pause très longue (enregistrement non coupé)
###	partie non transcrite (enregistrement non coupé)
$$$	coupure de l'enregistrement
=	liaison non-standard remarquable
#	absence remarquable de liaison
/..., .../	hésitations entre transcriptions
(...)	variantes graphiques indécidables
...-	amorces
*	syllabe incompréhensible
***	suite de syllabes incompréhensible
"..."	titres
...	nom propre ou suite de chiffres anonymisés

Tableau 1. Résumé des symboles de transcription

2. Locuteurs et tours de parole

2.1. Identification des locuteurs

Les locuteurs sont désignés par L1, L2, etc. *dans l'ordre de prise parole dans la partie transcrite.* Ils apparaissent en début de chaque tour de parole, séparés du texte par une tabulation.

L1 est-ce que je peux vous demander votre âge
L2 vingt-deux ans

Attention : (1) ne pas mettre d'espace ou tabulation avant la marque du locuteur ; (2) ne pas mettre d'espace entre le L et le numéro ; (3) toujours utiliser un L majuscule.

Dans le cas où il n'est pas possible de décider quel est le locuteur qui parle parmi L1, L2, etc. on utilisera la notation LX (attention L et X majuscules) :

L1 en ligne ça veut dire sur Internet
LX on line

Attention : LX ne signifie pas locuteur inconnu. Tout locuteur, même s'il ne prononce qu'un mot doit avoir un identificateur propre du type L1, L2, etc.

2.2. Chevauchements de parole

Les chevauchements de parole sont notés au moyen de chevrons < >. Le chevron ouvrant marque le début du chevauchement, le chevron fermant en marque la fin. Exemple :

L1 de me penser à penser à trouver un travail + < et l'imprimerie
L2 vous aviez à peu près > quel âge

Les chevauchements peuvent mettre en jeu plus de deux locuteurs :

L1 tu veux < que je te chauffe un peu le café pupuce
L3 oui je veux bien
L2 tu vois un peu hein c'est > + donc il faut faut s'intéresser à
L1 je vais mettre dans mon bol et je le mets au < micro-onde
L3 oui
L2 au fonctionnement > au fonctionnement même

Laisser une espace avant et après les chevrons (sauf en début et fin de ligne).

3. Règles typographiques

Ce sont celles de l'écrit standard, à part l'absence de majuscule en début de phrase, cette notion n'ayant pas cours dans les transcriptions d'oral (voir manuels de typographie).

> Il faut apporter beaucoup de soin aux détails typographiques. Les incorrections entraînent de grosses difficultés pour les traitements automatiques des corpus.
>
> Important : en cas d'utilisation de logiciels de traitement de texte évolués (du type *MS Word*) ne pas utiliser de caractères spéciaux, d'espace insécable, de tirets particuliers (cadratin ou autre), de mise en exposant ou indice, d'enrichissement typographique (gras, italique, etc.) ou de mise en forme (changement de marges, etc.). Attention aux automatismes de ces logiciels qui ont tendance à rajouter des majuscules en début de phrase, changer le type des guillemets ou des tirets, etc. Désactiver toutes les options ou mieux, utiliser un logiciel moins « intelligent », du type *Wordpad*.

3.1. Majuscules

On utilise normalement les majuscules sur les noms propres, géographique, historique, les titres et sigles, et autres mots portant normalement une majuscule à l'écrit :

ils ont créé le tout de **A** à **Z**
on a une zone industrielle de l'**Auxerrois** qui est assez dynamique
d'autres Anglais qui appartenaient à l'**O.M.S.**

> Attention: on met une majuscule sur le nom d'un peuple (le Français est gourmand), mais pas sur celui de la langue (j'apprends le français). De même, pas de majuscule sur les mois (en janvier et non en Janvier).

On utilisera les majuscules accentuées :

normalement "L'**Énéide**" est dans le programme de français
et alors j'ai appelé **Éléonore**

3.2. Sigles et acronymes

Dans la transcription, les sigles sont ponctués quand les lettres ont été prononcées isolément (S.N.C.F.), non ponctués lorsqu'il s'agit d'un acronyme prononcé comme un mot ordinaire (CROUS).

il peut apporter des **C.D.** à la **FNAC**
j'ai fini mon **DEUG** ça s'est bien passé

Le sens des sigles est précisé si besoin en commentaire lors de la première apparition (sigles peu courants uniquement) :

l'**A.E.E.** {sigle = Agence Européenne de l'Environnement}
actuellement il y a des négociations avec la **SANEF** {sigle = Société des Autoroutes du Nord Est de la France}

Les prononciations particulières doivent être notées (voir section Prononciation) :

une conférence de l'**I.E.E.E.** {sigle = Institute of Electrical and Electronics Engineers} {pron = [itRwaz2]}

3.3. Abréviations

En règle générale ne pas utiliser d'abréviations (titres honorifiques, unités, etc.), sauf dans les cas où la forme développée n'est normalement pas utilisée à l'écrit (etc.) :

Monsieur Chirac
en 300 **avant Jésus-Christ**
40 **kilomètres à l'heure**
douze ou treize **pour cent**
35 **degrés Celsius**

mais :

on nous ressort l'Inquisition les Dragonnades **etc.**

3.4. Lettres

Les lettres utilisées en tant que telles sont écrites en majuscules :

ils ont créé le tout de **A** à **Z**
un rang euh catégorie **B**

C'est le cas des mots épelés :

L1 fees {lang = anglais} c'est des c'est des comment est-ce qu'on dit c'est de la rémunération en fait + c'est de la parti- c'est c'est
L2 c'est un mot anglais
L1 oui c'est un mot anglais ouais **F E E S** +

3.5. Nombres et chiffres

Les nombres doivent respecter les normes habituelles de l'écrit. On écrit en lettres les nombres inférieurs à 10, les nombres employés comme substantifs, certaines expressions :

il pleut depuis **trois** jours
il y a **six** mois
il a raté sa **quatrième**
les années **cinquante**
la guerre de **Trente** ans

Dans les autres cas, écrire en chiffres (on sépare les tranches de mille par des espaces, sauf dans les dates) :

la lumière parcourt **300 000** kilomètres par seconde
il est né en **1986**

Respecter le format habituel des numéros de téléphone :

ce sera juste après le journal de dix-huit heures avec vos questions + au standard au **05 59 59 17 17**

3.6. Titres d'œuvres

Respecter l'écriture habituelle des titres d'œuvres littéraires, journaux, films etc., avec une majuscule sur le premier mot, et sur les noms propres internes. Isoler le titre du reste de la transcription par des guillemets droits.

on est allé regarder **"Matrix"**
le roman **"Vingt mille lieues sous les mers"**
j'ai revu **"Les vacances de Monsieur Hulot"** de Tati
je lis régulièrement **"Le Monde"**
je suis abonné au **"Monde"**

Attention : utiliser impérativement les guillemets droits. Pas d'espace entre les guillemets et le titre qu'ils encadrent.

3.7. Espaces

Ne jamais redoubler les espaces.

Ne pas laisser d'espace :

- entre l'amorce et le tiret (*un héli- hélicoptère*)
- entre la consonne de liaison remarquable et le signe = (*quatre =z= yeux*)
- entre les guillemets droits et le titre qu'il encadrent (*"Le Monde"*)
- avant la virgule de multi-transcription (*/était, a été/*)
- à l'intérieur des parenthèses de variantes (*on (n') a pas*)

Toujours laisser une espace :

- après le tiret d'amorce (*un héli- hélicoptère*)
- avant et après le signe de pause + (*il répond + à*)
- avant et après les signes d'interruption ///, ###, $$$
- après la virgule dans une double écoute (*/était, a été/*)
- avant et après les {...} et les [...]
- avant et après (n') (*on (n') a pas*)
- avant et après les chevrons de chevauchement (*L1 tu la trouves < bien, L2 très belle > oui*)
- avant et après # (de bons # amis)
- avant et après les *, ***, *P*, etc.

4. Orthographe

4.1. Accords non standard

Ne pas corriger, mais ajouter l'indication {sic}. Cette notation n'a aucun caractère de jugement normatif, mais indique simplement qu'il ne s'agit pas d'une erreur du transcripteur :

les conseils **nationals {sic}**
on gagne pas des sommes **exorbitants {sic}**

4.2. Élisions non réalisées

Ne pas corriger :

nous savions déjà par le "Le Guide du Routard" **que il** y avait quelques problèmes avec les enfants de de Djenné
elle est passée de quinze ouvrières à deux **parce que elle** pouvait plus euh les payer

4.3. Variantes morphologiques indécidables

En cas de variantes morphologiques non réalisées à l'oral, la transcription est indécidable et les alternatives sont notées entre parenthèses :

l'avantage de de l'I.U.T. euh c'est qu'il y a il y a pas de **spécialité(s)** euh **définie(s)**
quelques-unes à ne pas mettre entre toutes les mains comme **il(s) disai(en)t**
on **(n')** est pas quitte avec les gens

Sauf indication explicite du contraire, les adjectifs et participes se rapportant au pronom *on* restent au singulier, sauf réalisation phonique du pluriel. Exemple :

on est **parti** en Afrique de l'Ouest
on est **parti** tous les deux

mais :

on s'est **mises** dans de beaux draps

4.4. Onomatopées

Un certain nombre d'onomatopées sont codifiées et doivent être transcrites selon l'orthographe fournie :

ah, aïe, areu, atchoum, badaboum, baf, bah, bam, bang, bé, bêêê, beurk, bien, bing, boum, broum, cataclop, clap clap, coa coa, cocorico, coin coin, crac, croa croa, cuicui, ding, ding deng dong, ding dong, dring, eh, eh ben, eh bien, euh, flic flac, flip flop, frou frou, glouglou, glou glou, groin groin, grr, hé, hep, hi han, hip hip hip hourra, houla, hourra, hum, mêêê, meuh, mh, miam, miam miam, miaou, oh, O.K., ouah, ouah ouah, ouais, ouf, ouh, paf, pan, patatras, pchhh, pchit, pff, pif-paf, pin pon, pioupiou, plouf, pof, pouet, pouet pouet, pouf, psst, ron ron, schlaf, snif,

splaf, splatch, sss, tacatac, tagada, tchac, teuf teuf, tic tac, toc, tut tut, vlan, vroum, vrrr, wouah, zip.

Exemples:

hum hum je n'en suis pas si sûr
et alors **crac** le truc s'est cassé
et d'un coup j'entends **miaou miaou** au dessus de ma tête

Pour les onomatopées qui ne figurent pas dans cette liste, l'orthographe est libre.

4.5. Orthographe de mots inconnus

Lorsque l'orthographe est incertaine (cas de certains noms de marques, toponymes, etc.) on pourra utiliser une orthographe approximative *si elle est plausible,* en la faisant suivre de la mention {approx} (cas d'un mot unique) ou en l'englobant dans la séquence {début approx} ... {fin approx} (plusieurs mots) :

un certain **Dupont {approx}**
le lieu-dit **{début approx} Sous la Voivre {fin approx}**

Si aucune orthographe n'est plausible, on utilisera une marque de séquence incompréhensible *, *** ou une transcription phonétique entre [...] (voir plus bas).

Attention : il faut utiliser cette possibilité avec précaution, et ne pas inventer d'orthographes farfelues. S'il y a réelle difficulté de transcription, on utilisera la notation * ou ***, en ajoutant la prononciation en commentaire uniquement si elle a été perçue de façon certaine (voir plus bas, *Difficultés de transcription*).

4.6. Mots étrangers

Les mots étrangers reconnus par le transcripteur sont orthographiés conformément à la norme de la langue d'origine (majuscule aux noms allemands, etc.) ou dans un système de translittération standard (pour le russe, chinois, etc.). Lorsque le mot n'est pas largement adopté en français, ajouter une indication en commentaire du type {lang = ...} :

comment on dit **maze {lang = anglais}**

mais :

ils se sont garés sur le **parking** du supermarché
on devrait normalement avoir une + une séance de ce qu'ils appellent **debriefing** enfin

Lorsque plusieurs mots consécutifs sont dans une langue étrangère, utiliser la notation {début lang = ...} ... {fin langue}

eh ben peu à peu il devient **{debut lang = allemand} eine Sache {fin langue}**

Si le mot n'est pas reconnu, on utilisera la notation * ou *** (voir plus bas).

5. Prononciation

Les prononciations particulières sont notées entre crochets carrés [...] à l'aide de l'alphabet SAMPA (Annexe 3).

5.1. Prononciations déviantes

Le mot est parfaitement reconnaissable, mais sa prononciation est déviante (lapsus, prononciations non normatives, etc.). Dans ce cas, le mot est transcrit dans son orthographe habituelle et la prononciation est donnée en commentaire :

il se **passe {pron = [plas]}** plein de choses
les les **gens {pron = [Sa~]}**
pneumonie {pron = [pl2moni]}
aéroport {pron = [aReopOR]}

Si une prononciation s'applique à plusieurs mots, elle doit être encadrée par une balise ouvrante et une balise fermante {début pron = [...]} ... {fin pron}.

on est passé dans la boutique **{début pron = [tytifRi]} duty free {fin pron}**

5.2. Liaisons

On ne notera que les liaisons ou absences de liaisons réellement remarquables. Il ne faut pas noter les cas banals (*ils **sont à** Paris*).

On notera les liaisons remarquables en marquant la consonne entre deux signes = :

quatre **=z=** amis
donne-moi **=z=** en
on va écouter plutôt les vieux qui **=z=** ont vécu
le texte sur la vie financière + qui a été voté à mon avis a été **=t=** une sorte
je sais pas quand **=t=** euh ben
mon grand-père était **=t=** euh cordonnier
ça c'est le travail de certains euh **=z=** employés

L'absence d'une liaison obligatoire se notera avec le signe # :

c'étaient trois **#** amis

5.3. Prononciations courantes

Il ne faut pas indiquer la prononciation élidée (sans schwa) pour les clitiques (*je dis, tu le sais,* etc. prononcés [Zdi]). De même, il est inutile de noter les prononciations courantes de type *il y a* prononcé [ja], *tu sais* [tze], *tu as* [ta], *puis* [pi], *peut-être* [ptEt].

Lorsqu'une prononciation est à la fois remarquable et récurrente pour un locuteur, on ne la reportera pas dans la transcription, mais on la fera figurer en commentaire dans la fiche signalétique.

6. Phénomènes propres à l'oral

6.1. Amorces (mots inachevés)

Les amorces de mots sont notées par un tiret final (sans espace *avant* le tiret) :

une des rares **rou-** routes goudronnées
oui mais par **con-** bon au niveau euh euh au niveau social

Les notations *j'-* , *l'-*, etc. sont exclues. Utiliser *j-*, *l-* :

je me suis dit **j-** **j-** j'ai dit
l- l'association

6.2. Pauses et interruptions

Toutes les pauses devront être marquées à l'aide du signe +, même brèves.

mais + un jour de septembre 1937 + il me convoqua chez lui + il me dit tu sais + je suis gravement malade +

Les interruptions longues du discours (plusieurs secondes), le plus souvent liées à un événement particulier (le locuteur écrit au tableau, lancement d'un jingle à la radio, etc.), sont notées par une triple barre oblique ///. Une explication doit être fournie en commentaire :

/// {L1 se lève et va fermer une porte}
/// {musique}

Un certain nombre de mot-clés sont prédéfinis :

indicatif
jingle
musique
applaudissements

7. Difficultés de transcription

7.1. Mots incompréhensibles ou d'orthographe inconnue

Un mot ou une suite de mots incompréhensibles (inaudibles, dans une langue inconnue, etc.) seront notés * s'il s'agit d'une seule syllabe, *** s'il s'agit d'une suite de syllabes. Le cas échéant une transcription phonétique peut être fournie en commentaire, si elle est discernable (patronymes, toponymes ou mots étrangers, par exemple) :

le chemin était vraiment très long et que nous * la nuit a- arrivait
avec départ vers le puissant euh trois quart aile Frédéric * **{pron = [gaz]}**
je parle de Joseph *** je parle pas de de Jean-Pierre
vraiment dans la dans le coeur de l'Afrique où c'est euh à à *** **{pron = [basje]}**

Si la langue est inconnue, utiliser ??? pour le nom de la langue :

il lui a dit *** **{lang = ???}**

7.2. Multi-transcription

L'hésitation entre plusieurs séquences est notée entre barres obliques (sans espace après la barre ouvrante ni avant la barre fermante) :

après on a décoloré avec **/des, les/** acides
alors **/j'ai cherché, je cherchais/**

Lorsqu'il y a hésitation sur la présence d'une séquence sonore, on note une alternance avec 0 (attention, le chiffre zéro, pas la lettre O, ni le signe ensemble vide Ø) :

vous croyez qu'on a **/l', 0/** envie de ri- envie de rigoler

Il peut y avoir hésitation entre un morphème reconnu et une séquence incompréhensible. Dans ce cas on notera une alternance avec * ou *** :

bon ben de toute façon tu **/dois, */** avoir quand même l'expérience d'autres choses

7.3. Événements non linguistiques

Les rires, les bruits, sont signalés dans le texte entre accolades à l'endroit où ils se produisent s'ils sont ponctuels :

donc vous allez me parler de votre travail **{porte qui claque}**

ou bien sous forme de balises encadrantes si l'événement a une durée notable :

{début conversation de fond} ... **{fin conversation de fond}**

Si des événements sont récurrents dans un enregistrement (bruits, etc.), on ne les notera pas dans la transcription, mais en commentaire dans la fiche signalétique.

Un certain nombre de mot-clés sont prédéfinis[1] (les parenthèses indiquent des abréviations possibles) :

[1] Ces mots-clés sont les valeurs par défaut du logiciel *Transcriber*.

Locuteur :

respiration (r)	bruit de bouche (bb)	rire
inspiration (i)	bruit de gorge (bg)	sifflement (sif)
expiration (e)	toux (tx)	
reniflement (n)	râclement	
souffle (pf)	éternuement	

Auditoire ou autres locuteurs :

rires
rire(s) en fond
toux en fond
applaudissements
conversation de fond

Bruits divers :

bruit indéterminé (b)
froissement de papiers (pap)
souffle électrique (shh)
bruits micro (mic)
indicatif
jingle
musique

8. Anonymisation

Dans les cas où les noms propres doivent être supprimés on respectera le codage suivant : *P* = patronyme ; *T* = toponyme ; *S* = marque commerciale, nom de société.

il me dit Monsieur ***P*** il y a pas de problème
et cette entreprise donc euh germano-française ***S***

S'il y a plusieurs noms propres de la même catégorie dans la même transcription, chacun reçoit une numérotation : *P1*, *P2*, etc. :

des gens comme monsieur ***P1*** et monsieur ***P2*** madame ***P2***
ce qui serait marrant c'est que tu ailles demander au propriétaire du ***S1*** là
tu vois le bar à vin

Lorsque des suites de chiffres (numéros de téléphone, numéros de rue, numéro de sécurité sociale, etc.) doivent être anonymisées, on utilisera le codage *C* :

rappelez-moi au ***C***
il habitait au ***C*** des Champs-Elysées

9. Relation avec l'enregistrement

9.1. Parties non transcrites

Une partie comportant de la parole non transcrite (mais non coupée sur l'enregistrement) doit être notée par ###. La raison doit être donnée en commentaire :

{longue déclaration en basque}

Un certain nombre de mot-clés sont prédéfinis :

inintelligible
faible
très faible
voix superposées

9.2. Coupures de l'enregistrement

Une coupure de l'enregistrement doit être notée par $$$. Un commentaire peut être ajouté si l'on connaît la raison de la coupure, et/ou sa durée :

$$$ {l'invité joue du piano}
$$$ {longue séquence musicale, durée = 13 min.}

Un certain nombre de mot-clés sont prédéfinis :

indicatif
jingle
musique
applaudissements

Annexe 3. L'alphabet SAMPA[2]

Consonnes

	Symbole	Exemple	Transcription
Plosives	p	pont	po~
	b	bon	bo~
	t	temps	ta~
	d	dans	da~
	k	quand	ka~
	g	gant	ga~
Fricatives	f	femme	fam
	v	vent	va~
	s	sans	sa~
	z	zone	zon
	S	champ	Sa~
	Z	gens	Za~
Nasales	m	mont	mo~
	n	nom	no~
	J	oignon	oJo~
	N	camping	ka~piN
Liquides	l	long	lo~
	R	rond	Ro~
Semi-consonnes	w	coin	kwe~
	H	juin	ZHe~
	j	pierre	pjER

[2] http://www.phon.ucl.ac.uk/home/sampa/home.htm

Voyelles

	Symbole	Exemple	Transcription
Orales	i	si	si
	e	ses	se
	E	seize	sEz
	a	patte	pat
	A	pâte	pAt
	O	comme	kOm
	o	gros	gRo
	u	doux	du
	y	du	dy
	2	deux	d2
	9	neuf	n9f
	@	justement	Zyst@ma~
Nasales	e~	vin	ve~
	a~	vent	va~
	o~	bon	bo~
	9~	brun	bR9~
	E/	= e ou E	
Indéterminées	A/	= a ou A	
	&/	= 2 ou 9	
	O/	= o ou O	
	U~/	= e~ ou 9~	