



DGLFLF

Paul CAPPEAU & Magali SELJIDO

Paul.Cappeau@univ-poitiers.fr

magali.sejido@wanadoo.fr

Les corpus oraux en français

(inventaire 2005 v.1.0)

Table des matières

INTRODUCTION	3
CONSTAT	3
BUTS POURSUIVIS	4
CONSTITUTION DE L'INVENTAIRE ET PROLONGEMENTS	4
LES CRITÈRES UTILISÉS	5
LE NOM DU CORPUS	5
RESPONSABLE, OBJECTIFS, TYPES D'UTILISATION	6
RESPONSABLE	6
OBJECTIFS	6
TAILLE	9
CLASSEMENT EN DURÉE	9
CLASSEMENT EN NOMBRE DE MOTS	9
PÉRIODE DE CONSTITUTION / DATE D'ENREGISTREMENT	10
LE CONTENU (NON MÉDIA / MÉDIA)	11
EN HEURES	11
EN NOMBRE DE MOTS	11
EN NOMBRE D'ENREGISTREMENTS	12
LA POLITIQUE	13
DES ÉMISSIONS DE RADIO	13
DISPONIBILITÉ	13
BILAN PROVISOIRE	13
INVENTAIRE EXHAUSTIF	14



Introduction

Le champ ouvert par la constitution et l'exploitation de corpus oraux ouvre de nombreux enjeux (scientifiques, patrimoniaux, politiques, etc.) et paraît promis à un bel avenir économique : commandes des laboratoires d'analyse automatique du langage et des recherches en intelligence artificielle, commande des compagnies de téléphone, des traducteurs-interprètes, mise au point de serveurs vocaux "naturels", etc. Pour répondre à ces diverses demandes, il faut pouvoir disposer de corpus de données orales enregistrées et transcrites qui atteignent une grande taille (10 millions de mots serait l'idéal). Avant cela, il est utile d'avoir une meilleure connaissance de l'état des lieux et de dresser un inventaire des corpus oraux existants.

Constat

Depuis une vingtaine d'années, le développement des grands corpus de langue écrite et de langue parlée a considérablement modifié l'approche du langage. On y a vu parfois la naissance d'une nouvelle linguistique moderne. La constitution de grandes banques de données a déjà changé l'idée que nous nous faisons du patrimoine linguistique d'une nation et de sa diffusion. Toutes les études sur la langue en ont été renouvelées et on ne peut plus concevoir de recherches ni d'applications dans l'enseignement supérieur qui pourrait s'en passer. Les corpus de langue orale y jouent un rôle particulièrement important.

v Pour les **données écrites**, un gros effort a déjà été fait par la France qui a abouti à la constitution de la banque de données *Frantext*.

v Pour les **données orales**, les travaux d'autres pays (la Grande-Bretagne, le Portugal, l'Italie, notamment, pour rester dans le domaine européen) ont montré la nécessité de disposer de corpus d'importance : constitution de nouveaux outils de description de la langue en vue de l'enseignement (en tant que langue maternelle et en tant que langue étrangère) ; constitution de banques de données pour servir de comparaison dans le domaine de l'acquisition du langage, des troubles du langage, de l'évolution de la prononciation et du lexique, etc.

On sait qu'il existe quantité d'initiatives éparpillées, faites avec des conventions différentes rarement compatibles entre elles ; il est souvent difficile de savoir quelle en est la qualité et comment on peut y avoir accès. Il est donc apparu urgent de connaître la réalité des moyens existants et d'en dresser un **inventaire raisonné**.

La DGLFLF était l'organisme le mieux à même de conduire ce projet car elle a déjà organisé une sorte de "tutelle" sur la langue. Elle connaît une grande partie des projets d'équipes universitaires, avec lesquelles elle est en contact. Sa réputation lui permet d'organiser des expertises.

Buts poursuivis

Cet inventaire vise à lutter contre la dispersion actuelle des informations et il a l'ambition de rendre visibles les ressources orales disponibles dans la communauté scientifique française et de faire ressortir les lacunes et les manques, afin de favoriser de nouvelles recherches. Ce projet pourrait permettre, par la suite, de constituer un réseau donnant accès à ces données jusque là "éparpillées" et non recensées.

On pourra mieux coordonner les projets futurs en connaissant avec précision ce qui est disponible et les manques apparaîtront avec plus de relief. Cet inventaire permettra aussi aux équipes de communiquer entre elles pour des échanges d'information et de travaux afin d'éliminer les doublons inévitables tant que cette information n'est pas partagée.

Constitution de l'inventaire et prolongements

L'inventaire présenté repose sur deux sources d'informations : les ressources accessibles sur le net ainsi qu'une enquête et des contacts qui ont pu être pris avec les différentes équipes ou individus sollicités¹. Les correspondants contactés ont, le plus souvent, témoigné d'un grand intérêt pour cette entreprise et certains sont impatients de pouvoir en disposer. Les chercheurs étrangers qui constituent des corpus de français parlé ont été particulièrement coopératifs.

Dans le document présenté en annexe (qui regroupe toutes les informations collectées), la source des données a été différenciée (en noir pour celles recueillies sur le net, en bleu pour celles obtenues après enquête).

Cette première version n'est encore qu'une **ébauche** : on risque notamment de remarquer surtout ce qui ne s'y trouve pas ! Il va de soi que les auteurs attendent des retours qui leur permettront de compléter ces données et de fournir une deuxième version bien plus satisfaisante². Par exemple :

- v on sait que certaines personnes contactées n'ont finalement pas répondu (manque de temps, difficulté à réunir les informations demandées, etc.) Peut-être la vision de ce premier résultat les incitera-t-elle à le faire ;
- v certaines informations recueillies manquent parfois de précision. Là encore la vision d'ensemble que permet cet inventaire aura probablement des vertus bénéfiques et permettra de proposer une version améliorée ;
- v la quasi totalité des corpus recensés ont été constitués par des linguistes. Il s'agit bien évidemment d'une lacune qui sera corrigée par la suite. On espère dans une prochaine version pouvoir intégrer des renseignements sur les corpus constitués dans d'autres disciplines (histoire, sociologie, ethnologie, etc.).

Les auteurs de cet inventaire peuvent être joints aux adresses suivantes :
Paul.Cappeau@univ-poitiers.fr
magali.seijido@wanadoo.fr

¹ Remerciements particuliers à D. Luzzati qui menait un projet parallèle et nous a communiqué de nombreuses informations.

² Il manque en particulier les corpus du LIMSI pour lesquels on attend des précisions.

2

Les critères utilisés

Pour faciliter le recensement, un certain nombre de rubriques avaient été proposées aux correspondants lors de l'enquête. C'est à partir de ces facteurs qu'a été dressé le tableau final qui figure en annexe³.

Le nom du corpus

On reprendra ici la définition du terme "corpus" que l'on trouve dans le *Guide des bonnes pratiques* (2005) édité par la DGLFLF, c'est-à-dire *une collection ordonnée d'enregistrements de productions linguistiques orales et multimodales*. C'est l'entrée que nous avons privilégiée car c'est souvent le moyen le plus pratique de renvoyer aux données utilisées.

La désignation des corpus que nous avons recensés se fait sur des critères disparates qui peuvent renvoyer :

- v au contenu, ex : AIR France porte sur la réservation de billets d'Air France ;
- v au nom du chercheur responsable de ce corpus, ex : ALLAIRE renvoie au corpus constitué par Suzanne Allaire ;
- v à la ville où ont été effectués les enregistrements, ex : ANGERS est constitué d'enregistrements effectués à Angers ;
- v à un sigle qui renvoie à un organisme institutionnel, ex CREDIF.

Parfois les dénominations peuvent emprunter à deux de ces rubriques (ex : PARIS 3 EA 1483). Lorsque le corpus n'avait pas de nom officiel, nous lui avons attribué arbitrairement le nom de la ville universitaire où exerçait le chercheur responsable du corpus (par exemple : GRENOBLE, LYON II) suivi généralement d'une précision portant sur le nom du chercheur responsable ou de l'équipe responsable du corpus.

Il est important que chaque corpus possède un nom et un seul pour éviter les problèmes de doublons (on en a d'ailleurs signalé quelques-uns dans l'inventaire). A l'avenir, l'idéal serait de disposer d'une codification qui permettrait de renvoyer de façon univoque à chacun des corpus recensés. Mais cette phase de standardisation est probablement prématurée. La solution provisoire que nous avons retenue est une esquisse de ce qui pourrait être envisagé.

Cet inventaire permet de retrouver quelques corpus que l'on peut qualifier d'historiques. Il s'agit des corpus suivants :

NOM du Corpus	Date	localisation
ARCHIVES DE LA PAROLE	1912...	BNF

³ Nous n'avons retenu que les corpus en français et avons pour l'instant laissé à l'écart les corpus d'autres langues parlées en France. Ainsi le très corpus THESOC (en occitan) ne figure pas dans cet inventaire. Merci à Jean-Philippe Dalbera (Université de Nice) qui nous a communiqué toutes les informations demandées sur ce corpus de 2000 h.

FRANÇAIS FONDAMENTAL	1951-1953	Toulouse
Corpus d'ORLEANS	1966-1970	ESLO et divers sites étrangers

Tableau 1 – Les corpus “historiques”

v **Les Archives de la parole**

Créées sous l'impulsion de Ferdinand Brunot en 1911, les *Archives de la parole* bénéficient des nouveautés techniques de l'époque (comme le phonogramme) qui permettent d'enregistrer et de conserver des manifestations de la langue parlée. Brunot lui-même enquête sur le terrain (dans les années 1912-1913) pour enregistrer notamment “la parole au timbre juste, au rythme impeccable, à l'accent pur” ainsi que les patois et dialectes.

Pour plus de précisions, consulter le site de la BNF :

<http://gallica.bnf.fr/ArchivesParole/>

v **Le Français fondamental**

Ce corpus a été réalisé dans les années 1951-1955. La partie orale comporte 275 enregistrements (qui correspondent à un peu plus de 300 000 mots transcrits). Elle a été complétée par des textes écrits (journaux). Seule subsiste la transcription, les bandes ont en effet été détruites.

Pour plus de précisions :

Gougenheim, G., Michea, R., Rivenc, P., Sauvageot, A. (1956). *L'élaboration du français fondamental*, Paris, Didier

v **Le Corpus d'Orléans.**

Il a été constitué à la fin des années 60 auprès de locuteurs vivant à Orléans. Il comporte 144 entretiens pour un total de 300 heures. On peut en trouver des parties dans les corpus suivants de l'inventaire : BELC, ESLO, ELILAP, GULICH, AMSTERDAM. Une nouvelle transcription est en cours de réalisation (cf. O. Baude).

Pour une présentation plus détaillée de ce corpus :

Bergougnieux, Gabriel. (1996). “Etude Socio-Linguistique sur Orléans (1966-1970)”, *Revue Française de Linguistique Appliquée*, I-2, 87-88.

Responsable, objectifs, types d'utilisation

Responsable

Le responsable du corpus est, en général, le responsable institutionnel du corpus. Il s'agit soit de la personne physique qui peut être contactée et qui a participé à la constitution et / ou à l'édition du corpus, soit d'une institution ou d'une équipe de recherches (GARS, DELIC par exemple). *Le guide des bonnes pratiques* (2005) permettra de mieux préciser la notion d'auteur lorsqu'elle s'applique à des corpus oraux.

Cette information est complétée par les coordonnées (adresse électronique, adresse postale) de la personne à contacter pour disposer d'informations supplémentaires.

Objectifs

Les corpus sont habituellement constitués en vue d'une utilisation précise (étude syntaxique, enseignement, étude des interactions, etc.) et cet objectif n'est, bien évidemment, pas sans

Inventaire des corpus oraux (2005 - 1)

incidence sur le corpus lui-même. La forme des échanges, la durée des enregistrements ou la source des données (média ou non média) sont les principales variables affectées.

Les objectifs peuvent être indiqués d'une façon plus ou moins précise. On présente en trois tableaux distincts (les corpus localisés en France, les corpus localisés à l'étranger, les projets de corpus) les différents objectifs déclarés⁴ :

Types d'études	NOM des Corpus
Sociolinguistique	BRANCA-PARIS3, PARIS-VIII, PARIS-X, STRASBOURG-TABOURET-KELLERT,
Psycholinguistique	CHLOE , GRENOUILLE-KERN, NIMH, SPENCER, WEIL
Psychologie cognitive	CREPCO
Langage enfantin	CREDIF, NANCY-CANUT, TOULOUSE-3,
Didactique, enseignement du français	BELC, CRELEF, ESLO, FRANÇAIS FONDAMENTAL, GRENOBLE-1 & 2, PARIS-V, RADIO-FRANCE, STRASBOURG-TABOURET-KELLERT,
Phonologie, Prosodie ...	C-ORAL-ROM, IMPLANTS COCHLEAIRES, LL (et KP), Lyon III, MALECOT, NANCY, NIMH, PARIS 3 EA 1483, PASSY, SPENCER, THESOC
Morphosyntaxe	CAFE, CLER, CORPAIX-2, CRFP-1, DELIC, GARS, GRE, HP, NANCY-DEBAISIEUX, NIMH, PARIS 3 EA 1483, PERPIGNAN, POI, SPENCER, THESOC, TOULOUSE-VERGELY-PREVOT, TOULOUSE-DUVIGNAU
Pragmatique	C-ORAL-ROM, OZKAN, TOULOUSE-VERGELY-PREVOT
Les interactions	LYON II-GRIC, LYON II-CLAPI, PARIS 3 EA 1483,
Analyse du discours	CAFE, MONTPELLIER, NICE-CHAUVIN, TOULOUSE-4,
Sémantique	TOULOUSE-VERGELY-PREVOT, TOULOUSE-DUVIGNAU
autres	Interprétation simultanée : LEDERER La subordination : ALLAIRE Diachronie : THESOC,

⁴ Un certain nombre de corpus dont les objectifs n'ont pas été précisés ne figurent pas dans ces tableaux.

Lexique : TOULOUSE-1
Pour l'institut de géographie : ANGERS

Tableau 2- Objectifs des corpus localisés en France

On indique à part les corpus à visée dialectale qui n'ont pas été recensés de façon systématique. Outre les corpus cités dans l'inventaire, il faudrait inclure par exemple THESOC (cf. note 2) :

Dialectologie, parlars régionaux	ALMURA, FOSSAT, GRENOBLE-MAILLARD, LYON-III, NICE-MELLET, STRASBOURG-TABOURET-KELLERT
---	---

Types d'études	Pays	Nom du Corpus
Sociolinguistique	Belgique Suède Danemark Royaume-Uni	ELILAP (corpus d'Orléans), VALIBEL FPM, UPPSALA-LINDQVIST HANSEN BRISTOL, COVENEY, ESLO-ORLEANS
Phonologie	Belgique Suède Etats-Unis	VALIBEL FPM, UPPSALA-LINDQVIST VIOLIN-WIGENT
Régionalismes	Belgique	VALIBEL
Analyses conversationnelles	Belgique Suède Allemagne	VALIBEL FPM GULICH, KOTSCHI, BIELEFELD
Sémantique	Suède	FPM,
Enseignement	Suède Danemark Royaume-Uni	UPPSALA-LINDQVIST CAEN BRISTOL, FLLOC LABEAU, VOIX D'Auvergne
Syntaxe	Suède	FPM, UPPSALA-LINDQVIST, SOUTHAMPTON,
Atlas linguistique	Allemagne	GOEBL

Tableau 3- Objectifs des corpus localisés à l'étranger

Types d'études	Pays	Nom du Corpus
Phonologie	France	PFC
TAL	France	OTG
Etudes linguistiques diverses	France Belgique	ACI, CRFP-2 BRUXELLES-DELVENNE...
Analyse de conversations Interactions	Suisse Suède	BALE FPM
Représentation sociales	Suisse	NEUCHATEL-PY
Sociolinguistique	Suisse	NEUCHATEL-MAITRE-WILD
Acquisition	Suède	INTERFRA

Tableau 4- Objectifs des corpus en cours de constitution

Taille

Il n'y a pas d'uniformisation dans les unités qui permettent d'indiquer la taille d'un corpus. On trouve dans l'inventaire diverses sortes d'indications qui dépendent en partie de la date à laquelle le corpus a été réalisé. Ainsi pour des corpus anciens, on trouve parfois le nombre de pages de la version éditée (ce qui laisse supposer que le corpus n'existe pas sous une forme électronique). Les trois principales unités utilisées sont : la durée des enregistrements en heures, la longueur de la transcription en nombre de mots et parfois simplement le nombre d'enregistrements. On trouvera plus loin (en 2.5) un classement détaillé des corpus selon ces trois paramètres.

Cette disparité rend, pour l'instant, difficile un décompte global des corpus oraux recensés dans cet inventaire. On maintient, dans cette présentation, les deux types d'entrées les plus significatifs (la durée, le nombre de mots).

Classement en durée

L'indication de durée peut parfois être trompeuse puisqu'elle renvoie à la durée des enregistrements mais ne garantit pas toujours que la totalité soit transcrite. Il est tout de même intéressant de faire ressortir les plus gros corpus en France et à l'étranger.

NOM du Corpus	Durée des enregistrements	Date
CLAPI	600 h	Non clos
ESLO (Corpus d'Orléans)	350 h	1968-1969

Tableau 5 – Les corpus français les plus volumineux (en temps)

On peut aussi renvoyer à deux sites français qui permettent, selon des modalités propres, d'accéder à des enregistrements de français parlé (dont une grande partie sont d'origine audiovisuelle) qui pourraient par la suite devenir des transcriptions :

- site de la BNF : www.bnf.fr
- site de l'INA : www.ina.fr

NOM du Corpus	Durée des enregistrements	Pays
ELILAP	500 h	Belgique
VALIBEL	373 h	Belgique
MITCHELL	200 h	Grande-Bretagne

Tableau 6 – Les corpus à l'étranger les plus volumineux (en temps)

Classement en nombre de mots

Lorsque la taille du corpus est précisée en nombre de mots, cela permet de savoir qu'il s'agit de transcriptions en format électronique. Cette indication est rarement fournie, peut-être parce qu'il s'agit d'une unité de mesure plus récente (liée à l'informatisation) et qui n'est pas jugée pertinente dans certaines études. Voici à titre d'illustration, les trois plus gros corpus pour lesquels le nombre de mots était précisé (ce qui ne préjuge en rien de leur importance par rapport aux autres corpus) :

NOM du Corpus	Nombre de mots	Pays
ESLO / ORLEANS	4 500 000 mots	Royaume Uni
CORPAIX-2	1 702 000 mots	France
DELIC	1 460 000 mots	France

Inventaire des corpus oraux (2005 - 1)

ELILAP	1 000 000 mots	Belgique
GRE	500 000 mots	France
RINDLER-SCHJERVE	500 000 mots	Autriche

Tableau 7 – Les corpus transcrits les plus volumineux

On peut s'appuyer sur une équivalence assez sommaire entre nombre de mots et durée pour avoir une meilleure idée de la taille de l'ensemble des corpus situés en France. On considère qu'une minute correspond à 150 / 180 mots. On prendra donc une moyenne de 165 mots par minute ou encore 100 heures d'enregistrements représentent 1 millions de mots. On peut alors à partir des corpus déclarés en durée et en nombre de mots proposer le tableau suivant :

Pays	Equivalence en heures	Equivalence en nombre de mots
FRANCE	2 355 heures	23 550 000 mots
BELGIQUE	911 h.	9 110 000
SUISSE	25 h.	250 000
SUEDE	66 h 30	6 650 000
DANEMARK	29 h	290 000
FINLANDE	plus de 18 h	180 000
Pays-Bas	non identifiable	
ROYAUME-UNI	673 h	6 730 000
ALLEMAGNE	13 h 30	1 350 000
AUTRICHE	50 h	500 000
ETATS-UNIS	3 h	30 000

Tableau 8 – Total par pays

Il faut préciser que le corpus d'Orléans qui est disponible sur plusieurs sites est totalisé plusieurs fois dans ce décompte global. Pour la Belgique, le corpus d'Orléans (ESLO / ORLEANS) remplit 315h sur les 911 identifiées.

Période de constitution / date d'enregistrement

Cette indication permet de distinguer les corpus clos et les corpus ouverts qui continuent à être enrichis. Les deux informations ne sont d'ailleurs pas équivalentes puisqu'on peut très bien inclure dans un corpus ouvert des données enregistrées il y a plusieurs années. Le tableau ci-après fournit le relevé des corpus ouverts (on n'a pas tenu compte des corpus déclarés comme "en projet" qui figurent dans l'inventaire global en section 2) :

Nom du Corpus	Date de création
ALMURA	2000
BRANCA-PARIS3	2002
CLER	1996
DELIC	2000
GRE	1996
NANCY-DEBAISIEUX	2000
POI	1993

Tableau 9 – Principaux corpus ouverts (en France)

Le contenu (non média / média)

On ne dispose pas d'une claire vision du contenu des corpus recensés. Ce serait d'ailleurs une indication utile pour avoir une meilleure vision des domaines bien représentés et des lacunes. Pour l'instant on se contentera d'une répartition selon la source d'enregistrement (non média / média). Dans ce dernier cas, on peut approfondir quelque peu les contenus en prenant en compte le type d'émissions enregistrées.

On fournit ci-après la répartition (à titre indicatif⁵) des corpus français entre non média et média à partir des indications fournies. Trois indicateurs ont été utilisés : le nombre d'heures des enregistrements, le nombre de mots des transcriptions et à défaut d'autres indications le nombre d'enregistrements. Dans ce dernier cas, la mesure est très vague puisqu'on ne maîtrise pas la longueur variable des enregistrements :

En heures

Non média		Média	
Taille	Nom du corpus	Taille	Nom du corpus
40 h	BANGE	20 h	ALLAIRE
6h 30	BELC	11h	BRANCA-PARIS3
3 h	CHLOE		
26 h	CLER		
16 h	C-ORAL-ROM	8 h	C-ORAL-ROM
350 h	ESLO		
60 h	FRANÇAIS FONDAMENTAL	15 h	FRANÇAIS FONDAMENTAL
1 h	GRENOBLE-1		
15 h	GRENOBLE-MAILLARD		
50 h	LEDERER	35 h	HP
14 h	LL & KP		
600 h	LYON II-CLAPI		
40 h	LYON II-RITTAUD		
20 h	LYON III		
2 h	MICROFUSEES		
29 h	NANCY		
45 h	NANCY-DEBAISIEUX	5 h	NANCY-DEBAISIEUX
3h 30	NICE		
456 h	NIMH		
100 h	PARIS VIII		
5 h	TOULOUSE-DUVIGNAU		

Tableau 10 – Répartition des corpus français (en non média / média) selon la durée

En nombre de mots

Non média		Média	
Taille	Nom du corpus	Taille	Nom du corpus
54 000	AIR FRANCE		

⁵ Lorsqu'un corpus contient des enregistrements mixtes (média et non média) la répartition est souvent approximative. On a considéré de façon arbitraire que 10 % était réservé aux média.

Inventaire des corpus oraux (2005 - 1)

7 000	CAFE		
1 500 000	CORPAIX-2	200 000	CORPAIX-2
1 300 000	DELIC	160 000	DELIC
500 000	GRE		
11 500	OZKAN		
118 726	TOULOUSE-VERGELY		

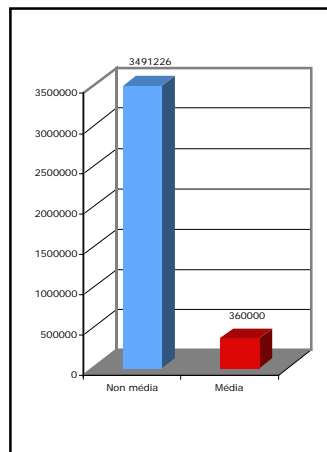
Tableau 11 – Répartition des corpus français (en non média / média) selon la taille

En nombre d'enregistrements

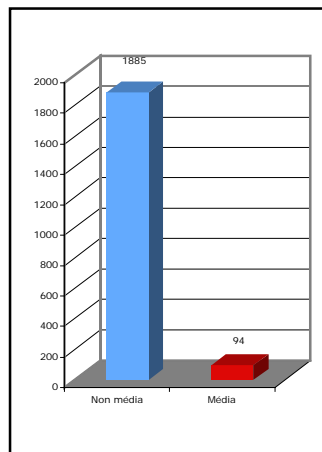
Non média		Média	
Taille	Nom du corpus	Taille	Nom du corpus
50	ALMURA		
120	CRFP-1	14	CRFP-1
140	GRENOUILLE-KERN		
50	MALECOT		
500	NANCY-CANUT		
180	PARIS 3 EA 1483	20	PARIS 3 EA 1483
3	PARIS VII		
150	PERPIGNAN		
1	PIC		
139	POI	11	POI
27	RENAULT		
320	SPENCER		
7	STRASBOURG TABOURET-KELLER		
120	WEIL		

Tableau 12 – Répartition des corpus français (en non média / média) en nombre d'enregistrements

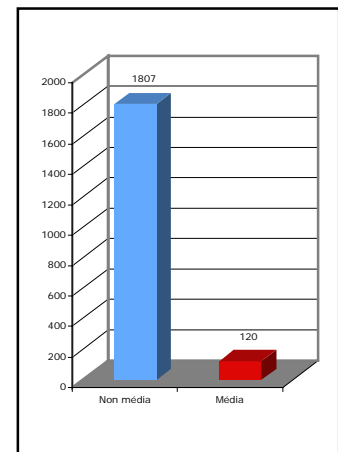
Les graphiques suivants permettent de synthétiser les données précédentes et de constater la part relativement faible des corpus provenant des média.



1. en nombre de mots



2. en durée



3. en nombre d'enregistrements

Pour les corpus média, une subdivision est envisageable selon le contenu :

La politique

NOM du Corpus	Contenu	Date
ALLAIRE	Débats politiques	1967
HP	Interviews / débats	1995...

Il faut ajouter à ce relevé deux corpus qui sont constitués de l'enregistrement de journaux télévisés (ce qui laisse supposer qu'ils contiendront des séquences liées à la politique) :

NOM du Corpus	Contenu
FPM (Suède)	Journaux télévisés
UPPSALA-LINDQVIST (Suède)	Journaux télévisés

Des émissions de radio

NOM du Corpus	Contenu
CRELEF	Non précisé
PARIS 3 EA1483	Radioscopie
LABEAU	Non précisé

Disponibilité

L'accès aux données déclarées à travers cet inventaire est très disparate. Cette information est fournie dans la colonne *consultation* du tableau général

- v de nombreux corpus sont consultables sur place (texte et son le plus souvent) ;
- v plusieurs corpus sont accessibles via internet (soit en totalité, soit pour partie), librement ou avec mot de passe ;
- v enfin de nombreux corpus ne sont pas (encore ?) communicables, soit parce qu'ils sont considérés comme corpus de travail d'une équipe, soit parce qu'ils existent sous une forme écrite seulement, soit parce qu'ils ne sont pas encore transcrits.



Bilan provisoire

Même si cet inventaire n'est pas complet, il est tout de même possible de dégager quelques enseignements des informations rassemblées :

- v les corpus oraux déclarés représentent une masse importante de données. Celles-ci sont toutefois difficiles à regrouper pour diverses raisons qui ont été évoquées :

- les corpus ont été constitués sur une période longue et donc ils se présentent sous des formats disparates (support papier uniquement, données informatisées) ;
 - il est parfois difficile de faire la part entre les données enregistrées et les données transcrites ;
 - les conventions de transcriptions ne sont pas homogènes. Cela ressort pour partie des objectifs différents qui ont présidé à la constitution des corpus. Cette diversité peut contrarier les échanges de données ;
 - l'accès aux données est très variable : certains corpus peuvent être consultés sur le net, d'autres ne sont accessibles qu'aux membres d'une équipe, ce qui rend quelque peu factice ou virtuel la vision d'un corpus globalisé.
- v on dispose de peu d'indications sur le contenu des corpus, ce qui rend là encore peu aisé un regroupement des données.

Pour l'instant on se trouve en présence d'un ensemble vaste de données mais cet ensemble apparaît à bien des égards hétérogène, disparate. Il semble donc que l'on soit encore un peu éloigné de l'objectif poursuivi dans plusieurs pays européens de constituer un corpus de plus de 10 000 000 de mots. La France constitue un cas à part.

Toutefois, il existe un fort intérêt pour les corpus oraux et il existe de nombreuses équipes qui ont un savoir faire. On peut alors espérer que des projets futurs sachent faire converger ces énergies.

4

Inventaire exhaustif

Le fichier Excel comporte un sommaire avec des liens hypertexte qui permettent d'accéder directement aux rubriques désignées.

Lorsque les corpus sont suffisamment nombreux, on a aussi utilisé une présentation avec un gestionnaire de liste qui permet un tri sur chacune des colonnes de l'inventaire. Il suffit pour cela de cliquer sur les petits triangles à droite de chaque colonne (cf. figure ci-dessous) et de choisir les options de tri qui apparaissent.

Nom du corpus	Responsable, objectifs
---------------	------------------------

[Voir le fichier](#)