

ÉLABORATION DU CORPUS EN FRANÇAIS PARLÉ AU CONGO

Edouard Ngamounsika,
edouard.ngamounsika@umng.cg
Université Marien Ngouabi

Résumé : *Cet article évoque la question de l'élaboration du corpus du français parlé recueilli en République du Congo (CFPC) en vue de servir aux analyses des différentes variétés linguistiques du français. Le CFPC permet d'étudier des usages authentiques de la langue française. L'article s'intéresse donc au problème de constitution du corpus, du choix des données, des conditions d'enregistrement, du choix de transcription et des problèmes juridiques.*

Mots clés : oral spontané, français parlé, Congo, variétés linguistiques

Summer: *This article discusses the development of the spoken French corpus collected in the Republic of Congo (CFPC) in order to analyze the different linguistic varieties of French. The CFPC makes it possible to study authentic uses of the French language. The article is therefore concerned with the problem of the constitution of the corpus, the choice of the data, the conditions of registration, the transcription choice and the legal problems.*

Keywords: spontaneous oral, spoken French, Congo, linguistic varieties

Introduction :

Dans le cadre de cette première journée scientifique du Centre de Recherches en Linguistique et langues orales (CERELLO), je me propose d'évoquer la question de la constitution du corpus notamment du corpus du français parlé au Congo (désormais CFPC)¹. Il est bon de rappeler que le mot corpus

¹ En effet, l'idée de Constituer le CFPC remonte à 2004 lorsque nous débutions notre thèse de doctorat portant sur *Le Français parlé en République du Congo : étude morphosyntaxique*. Au laboratoire du Groupe Aixois de Recherche en Syntaxe (GARS) au sein de l'université de Provence Aix-Marseille I, nous constatons la présence des corpus tels que Corpus d'Aix, le CFPP2000, le corpus d'Oxford, d'Orléans, etc., il manquait un corpus du français parlé africain.

renvoie à un ensemble de documents de contenus divers (textes, images, films, etc.) regroupés dans un objectif déterminé, généralement pour servir de matériau des études de linguistiques.

Notre objectif, ici, est de constituer un CFPC qui prendrait en compte toutes les variétés linguistiques congolaises. Comme le souligne, en effet, Elisabeth Delais-Roussarie (2002, p5) :

les corpus sont de plus en plus importants, et cela dans de nombreux domaines : traitement automatique de la parole, traitement automatique des langues naturelles, didactique des langues, linguistique descriptive, psycholinguistique, sociolinguistique, etc.

Nous pensons que les corpus de langue parlée sont des creusets féconds d'indices mettant efficacement et efficacement en exergue les éléments centraux et typiques d'une langue et dans une langue. Autrement dit, les corpus rendront compte des aspects normatifs des langues en leur structure et en leur langue sans tracer une frontière étanche avec leur *status*, c'est-à-dire les conditions d'utilisation desdites langues. Par où l'on voit qu'il est expressément nécessaire de travailler sur des enregistrements et des transcriptions authentiques. Par où l'on voit aussi, que l'approche hypothético-déductive, en la matière, montre des limites et des faiblesses. A ce propos, Claire Blanche-Benveniste (1996, p25) citant Sinclair confirme ce qui suit:

Quand on invente des exemples, on confond souvent l'exemple et l'explication, l'exemple étant construit précisément pour justifier et clarifier l'explication. On ne peut pas inventer ce qu'est l'usage ; on peut seulement l'enregistrer.

L'analyse d'un corpus permet d'étudier les usages authentiques de la langue française car on ne peut pas se permettre d'inventer les exemples de langue parlée en se fiant à l'intuition, qui est souvent trompeuse. L'article s'appuie sur le problème de constitution du corpus, du choix des données, des conditions d'enregistrement, du choix de transcription et des problèmes juridiques.

Depuis cette époque, nous avons poursuivi la constitution du CFPC avec l'aide de nos étudiants de Master et de thèse.

1. Problème de constitution du corpus

La constitution du corpus oral soulève deux problèmes : la réticence des locuteurs interviewés et le coût lié à son élaboration.

1.1. La peur des informateurs et de leur réticence

La constitution d'un corpus suscite toujours des problèmes de plusieurs ordres que nous allons énumérer. L'un des premiers problèmes rencontrés est celui de la peur des informateurs et de leur réticence. C'est le cas des étudiantes de Licence en science et techniques de communication (STC) de la Faculté des Lettres des sciences Humaines de l'Université Marien Ngouabi dans le Texte *dans la cour* (158, 3-17) quand l'une d'elles a vu le micro:

L4 monsieur il fallait d'abord demander notre X consentement

L1 ah bon

L3 il a dit qu'il

enregistrait par rapport

XX L4 XX

L1 au départ je vous ai présenté mon appareil

L3 il a dit il a dit qu'il enregistrait pour

L4 X X

L3 tu ne voyais pas donc – il dit j'enregistre pour

L1 mes recherches

L2 mes recherches sur le non- tout tout et quand il a appuyé

j'ai même vu qu'il était appuyé-

L1 n'ayez pas peur hein- on parle de tout sinon de rien

L3 vous n'avez pas peur

L1 vous n'avez rien à craindre- vous ne critiquez personne- ce que vous dites est vrai-

Voici la façon dont nous procédions. Nous priions les locuteurs d'accepter d'être enregistrés. Nous les instruisions sommairement de nos desseins «il s'agit d'une étude linguistique. » Ensuite, nous leur montrions le dictaphone et nous engagions la conversation. Comme pour les enregistrements du français fondamental, nous nous sommes « efforcés d'éviter les sujets d'ordre politique, religieux ou personnel qui auraient pu provoquer certaines réticences chez les témoins » (Gougenheim *et alii*, 1956 : 58).

1.2. Le coût lié à l'élaboration du corpus

Le deuxième problème est lié au fait que « les données de langue parlée, qui impliquent enquêtes, enregistrements et transcriptions, sont beaucoup plus "chères" que des données écrites » (Blanche-Benveniste, 2005 :76). Il faut reconnaître que l'élaboration du corpus est très coûteuse, en temps, en personnel et en matériel. Comme le soulignaient Queffélec et Niangouna (1990, p. 46)

Pour faire une étude scientifiquement fondée du français oral utilisé par la population congolaise, il aurait fallu procéder à la manière des enquêtes mises au point pour l'élaboration du français fondamental de G. Gougenheim et utilisées dans certaines descriptions des langues africaines: en aurait été nécessaires de nombreux enregistrements opérés si possible à l'insu des intéressés, en divers lieux (bar, marché, services publics, taxis, etc.) où le français est parlé, si possible auprès d'un échantillon représentatif de la population congolaise parlant français et ce, dans les diverses régions du pays. Un tel programme dépassait très largement nos possibilités. Un embryon de recherches allant dans cette direction a certes existé : des émissions de radio ou de télévision, censées représenter la norme locale orale ont été enregistrées (...)

Selon Ambroise Queffélec (1997, p.355) « constituer un corpus oral est effectivement onéreux en temps et en moyens» puisque chaque minute d'enregistrement, avant d'« être disponible sur le support informatique sous forme fiable, «coûte» au minimum une heure de travail, et ce pour du personnel déjà entraîné à ce genre» (Delcourt, Francard & Moreau 1993 :325). Nous pensons aussi avec Claire Blanche-Benveniste (1999, p.66) que

la limitation vient évidemment du temps nécessaire à l'établissement des transcriptions de langue parlée, qui varie considérablement selon les types de transcription adoptés. Cette étape, autrefois considérée comme mineure, et souvent confiée à de simples exécutants, est devenue un travail de spécialistes.

Le CFPC se voudrait un corpus congolais dont on retrouverait des enregistrements de tous les départements du Congo quoique de taille

limitée. Nous envisageons 5 à 10 heures de temps soit 120 heures d'enregistrement.

2. Choix des données

Notre objectif est d'observer et de décrire les usages authentiques de la langue parlée par les locuteurs congolais. C'est pour cela que nous n'avions pas de questionnaire préétabli à soumettre aux locuteurs pour réaliser une enquête sociolinguistique particulière. Celle-ci s'inscrirait dans la perspective de Knutsen (2007, p.110)

Le but de l'enquête sociolinguistique était dans un premier temps de répertorier l'opinion des informateurs sur les rapports entre les langues à Abidjan, leur choix de langue dans tel ou tel contexte, la place du français dans le plurilinguisme à Abidjan et l'évaluation des différentes pratiques langagières en français afin de déterminer dans un second temps l'appréciation de l'informateur sur les normes et leur légitimité.

Nous estimons que dans **l'enquête sociolinguistique**, c'est le linguiste qui oriente le locuteur en vue de vérifier ses intuitions. Alors qu'en réalité, la démarche devrait être inverse. Nous avons pensé que le corpus seul fait émerger les aspects à traiter. Nous adoptons la position de Martinet cité par Maurice Leroy (1980, pp.91-92) :

Il faut répéter, une fois de plus, que ce n'est à la langue de se conformer aux édits des linguistes, mais aux linguistes d'adapter leurs méthodes si elles ne rendent pleine justice à la langue étudiée.

Une fois le corpus constitué, nous constatons que certains critères se sont imposés d'eux-mêmes.

2.1. Le degré d'instruction

La pratique du français au Congo est devenue, en quelque sorte, généralisée. Nous pouvons dire que presque tous les Congolais ont une compétence en français. Le critère de degré d'instruction n'est, cependant, pas très fiable dans la mesure où certains locuteurs, quoique ayant un niveau

d'instruction bas, sont à mesure de réaliser des énoncés corrects et vice-versa. Le même paradoxe s'observe avec des locuteurs lettrés qui produisent des énoncés non-standard. Ce phénomène montre la fragilité des distinctions rigides auxquelles donne lieu un classement des locuteurs en fonction du nombre d'années de scolarité.

Niveaux d'instruction	Énoncés produits	
	Correct	Incorrect
Élevé	●	○
	○	●
Moyen	●	○
	○	●
Bas	●	○
	○	●

Légende

- Effectif
- Non-effectif

Tableau récapitulatif des énoncés produits par les usagers de la langue française au Congo

2.2. L'âge

L'âge de tous nos informateurs varie de 20 ans à 45 ans. Il s'agit donc de locuteurs exerçant un métier dans la vie sociale. Aucune distinction n'est faite entre les hommes et les femmes. Mais les femmes se sont souvent montrées très réticentes aux enregistrements. Cela a été aussi le cas de certains hommes qui croyaient qu'il s'agissait de la police secrète. Aussi craignent-ils de voir les hommes avec un micro.

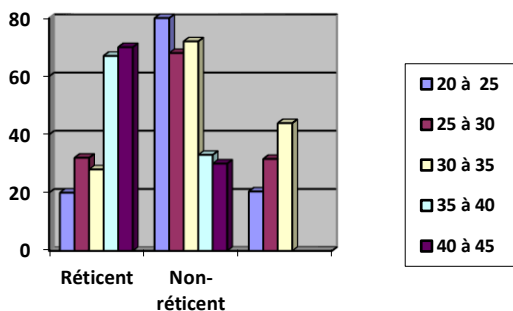


Diagramme des attitudes des enquêtés sur une population de 100 de 20 à 45 ans

3. Les conditions d'enregistrement

Nous avons recouru à deux types d'enregistrement. Le premier « sur le vif » qu'Hattiger (1981, p.36) définit ainsi : «l'enquête idéale aurait été celle qui eût permis de rassembler un corpus enregistré " sur le vif ", dans des situations de communication spontanée et sans que les locuteurs eussent conscience d'être enregistrés». Nous l'avons fait pour certains enregistrements avec micro invisible ou même avec micro visible. Lorsque nous demandons aux locuteurs de nous raconter quelque chose, au fur et à mesure, ils ne surveillaient plus leur performance langagière. Ils parlaient sans faire attention au dictaphone. Il n'est pas facile de demander aux locuteurs de préciser leur langue maternelle ou le lieu de leur résidence. Souvent certains enquêtés demandaient un intéressement à la fin.

3.1. L'enquêteur

Si les enregistrements de départ ont été réalisés par nos soins dans le cadre de notre thèse, les enregistrements actuels sont effectués par nos étudiants en Master et en Thèse.

3.2. Le lieu d'enregistrement

Les enregistrements ont eu lieu dans différents cadres comme les buvettes, les rues, la Faculté des Lettres, un restaurant ou une maison. Cette diversification de lieu d'enregistrements a favorisé le recours à plusieurs

situations. Cette liberté dans le choix du contexte d'enregistrement s'explique par la recherche de la diversité, de la spontanéité et du naturel. Les textes *Mbvuru* et *Dans la cour* ont comme support de véritables conversations regroupant plusieurs locutrices et locuteurs. Nous y remarquons une expression caractérisée par la spontanéité comme dans tout discours de conversation. Souvent, il leur était demandé de raconter comment ils avaient vécu des faits particuliers qui leur tenaient à cœur conformément à la recommandation de Labov (1976, p. 298) : « ...on peut également l'engager [l'informateur] dans des matières qui recréent des émotions violentes ressenties dans le passé, ou qui l'impliquent dans un contexte différent ».

3.3. La durée des enregistrements

Nos enregistrements durent au minimum 15mn, et varient entre 15 et 60 minutes. La durée de l'enregistrement est fonction de la disponibilité des locuteurs.

4. Transcriptions des données

Une fois le corpus constitué, la deuxième étape a consisté à le transcrire, c'est-à-dire à « donner une forme écrite à des enregistrements oraux » ce qui favorise ce que Queffélec (1997, p.362) appelle « l'inévitable perte d'information » car « transcrire c'est traduire et traduire c'est trahir ». Comme l'affirme Pierre Encrevé (1992, p104) :

Si on transcrit de l'oral, on en fait de l'écrit. Il faut absolument au contraire réussir à garder à l'oral toute son extraordinaire spécificité, toutes ces marques qui ne s'écrivent pas, même dans des alphabets phonétiques les plus complets [...] L'oral écrit, ce n'est plus de l'oral. Il a perdu son corps, il n'a plus que ce squelette nécessaire mais insuffisant qu'est l'écriture. Tout ce qui est le corps, comme le corps d'un humain, ses aspects spécifiques, ne peut pas être appréhendé une fois qu'on a transcrit.

Conscient de cette perte, le chercheur est obligé de transcrire ses enregistrements car « l'étude de l'oral sans transcription est extrêmement difficile, impossible même » (Chevalier, 1992, p.104). Queffélec pense, d'ailleurs, que le mode de transcription choisi doit être adapté aux

objectifs du chercheur qui ne sont pas les mêmes selon que ce dernier est phonéticien, morpho- syntacticien ou lexicologue.

4.1. Fidélité et lisibilité

Si, comme le souligne Poplack (1984, p. 24), « l'objectif d'une transcription est d'obtenir un rendu fidèle de ce qui a été dit, en respectant les omissions, les réinsertions de mots fonctionnels, etc. », le grammairien, peut se contenter d'outils de transposition moins sophistiqués. Il est, en cette activité, suivant Blanche-Benveniste & Jeanjean (1986, p.115) « tiraillé entre deux exigences la *fidélité* à la chose parlée et la *lisibilité* de son rendu par écrit». La fidélité à la parlure congolaise, en tenant compte des amorces, des énoncés, des écarts, des emprunts aux langues du substrat est l'un des soucis du transcripateur. Nous avons tenté de transcrire le FPC tel qu'il est employé par ses locuteurs ; conscients du fait qu'il est impossible de tout transcrire. Il y a inévitablement une perte d'information.

Déjà Blanche-Benveniste et Jeanjean (1987, p.113) s'interrogeaient : « peut-on vraiment "tout" garder ? Peut-on réellement être fidèle à la production orale en donnant une transcription écrite ? N'y a-t-il pas, dans toute transcription de l'oral, un appauvrissement des données ? ». La fidélité au français parlé africain en contexte de contact de langues est impossible. Il faudrait certainement transcrire la langue telle qu'elle est pratiquée : les tons, les bribes, la prosodie, bref tous les aspects caractéristiques d'une parlure africaine. La fidélité à la transcription est fonction des attentes et/ou du champ disciplinaire du chercheur. Nous nous situons dans une perspective morpho-syntacticienne.

La notion de **la lisibilité de transcription** dépend des objectifs du chercheur. Nous avons choisi une transcription non surchargée qui ne situe pas la qualité de la voix, les gestes, les mimiques, les rythmes. D'ailleurs comme le souligne Dérive (1975, p.53) « transposer un énoncé parlé en un énoncé écrit ne signifie pas lui faire perdre sur le papier le charme particulier que lui conférait son oralité, en l'habillant et en le sophistiquant.»

4.2 Les différentes transcriptions

La visée de l'étude : linguistique, sociolinguistique, ethnolinguistique, phonétique, prosodique, morphosyntaxique, etc. permet au linguiste de choisir sa transcription. Il existe deux modes principaux de transcriptions : phonétique et orthographique.

4.2.1. La transcription phonétique

Cette transcription convient aux phonéticiens. Selon Ambroise Queffélec (1997, pp.362-363)

pour un chercheur qui se propose de travailler sur la prosodie (définie comme l'ensemble des variations d'intensité, de fréquence fondamentale de la voix et de durée segmentale), l'enregistrement des données doit respecter des protocoles très précis et la transcription puis l'analyse des données nécessite des "stations informatiques avec des outils de traitement du signal de parole" (...) pour la transcription des français oraux d'Afrique, les transcriptions par les chercheurs dépendent de la variété de français enregistrée : pour les français basilectaux, la transcription phonétique (avec les normes de l'A.P.I) paraît s'imposer.

La transcription phonétique est rébarbative pour une analyse morphosyntaxique. Elle ne rend pas compte des phénomènes intonatifs et prosodiques jugés pourtant essentiels par Pierre Léon (1977, p:9) « pour faire une grammaire de l'oral. » La seconde difficulté est celle liée à l'exploitation informatique « pour les unités supérieures au phonème comme le mot ou syntagme qui intéresse le grammairien. » (Queffélec, *ibid.*) Il est difficile de lemmatiser ce corpus. Nous avons retenu la transcription phonétique pour des énoncés en langue bantou. Par exemple le texte *Un amour déçu* comporte des séquences transcrites comme :

yādi vāda ka na diboko yaku zengana

Nous trouvons, certes, des transcriptions phonétiques accompagnées d'une version orthographique très lourde quoiqu'elle soit une solution idoine pour le profane. Le GARS réserve cet usage aux productions marginales ce qui n'est pas le cas en FPC.

4.2.2. La transcription de l'intonation

La transcription de l'intonation est difficile d'usage car « pour noter l'intonation encore faut-il être spécialiste, et, à l'heure actuelle il semble bien qu'on ne dispose pas, parmi les spécialistes, d'une notation systématique qu'on pourrait appliquer à des corpus de français parlé » souligne Blanche-Benveniste et Jeanjean (1987, p129).

4.2.3. La transcription orthographique

La transcription de type orthographique est celle que nous avons retenue. Dans le numéro 18 de *Recherche sur le Français Parlé*, l'Équipe de DELIC (2004, p 22) en fait la synthèse suivante :

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 travers de certaines difficultés de transposition liées à des problèmes d'interprétation.

La transcription orthographique a le mérite de faciliter la lecture et le traitement informatique de même qu'elle fournit « des formes signifiantes, morphèmes que nous connaissons bien, sans nous préoccuper de leur réalisation phonique » (Blanche-Benveniste, 2005 : 48). L'absence de ponctuation ne pose en aucun cas un problème dans la mesure où la « ponctuation est un enjeu important, qui entraîne des décisions sur la segmentation des productions. Mettre des points c'est délimiter des phrases » précise Blanche-Benveniste (2005, p. 49). Les énoncés sont aménagés pour faciliter leur exploitation.

Nous avons donc adopté la transcription orthographique. La transcription phonétique n'a été utilisée que pour transcrire les énoncés en langue congolaise. Nous ne partageons pas, par conséquent, le point de vue de Knutsen (2007, p.88) qui pense que l'analyse morpho-syntaxique devrait opter pour la transcription phonétique tandis que l'étude sociolinguistique serait réservée à la transcription orthographique.

L'analyse (détaillée) de la structure morpho-syntaxique requiert dans la plupart des cas une transcription phonémique, tandis que l'analyse du discours ou l'analyse sociolinguistique se fait généralement à partir de la transcription orthographique. Alors que la transcription phonémique alourdit la lisibilité, surtout dans les cas où les données sont abondantes, la transcription orthographique présente l'inconvénient de faire disparaître des éléments d'information précieux, et parfois cruciaux pour la bonne interprétation structurelle et sémantique, étant donné son niveau d'abstraction élevé.

Ce choix de Knutsen (*ibid.*), s'explique par le malaise que suscite le français africain :

Un autre aspect souvent mis en avant, surtout en ce qui concerne la représentation des discours africains, est l'aspect " stigmatisant " de celle-ci. Étant donné la péjoration sociale du terme "petit-nègre", défini par le *Petit Robert* comme un " français incorrect et sommaire parlé par les indigènes des colonies, par extension mauvais français", les linguistes sont à cet égard sensibles à l'effet créé par la transcription. Cependant, ils ne sont pas d'accord sur la démarche à adopter pour échapper à cette stigmatisation.

Knutsen rapporte, pour évaluer l'intérêt mais aussi les inconvénients des deux modes de transcriptions, deux citations. Pour certains africanistes, la transcription de type orthographique ne serait pas adéquate, en particulier pour des corpus de non scolarisés comme ceux d'Hattiger (1981, p.45) qui a défendu « quand le besoin s'en fait sentir, la notation des tons montants ou descendants ». Mariette-Crespo (cité par Queffélec), au terme d'une longue discussion théorique, a aussi opté pour une transcription phonétique « intermédiaire » privilégiant l'allophone au phonème. A la lecture des travaux du GARS, nous pensons que les morpho-syntacticiens, les lexicographes « préfèrent des systèmes de transcription qui respectent pour l'essentiel le code orthographique français » (Queffélec, 1987, p.363) qui n'est donc pas seulement l'apanage des sociolinguistes.

5.3. Les difficultés de transcription

Une fois le type de transcription adopté, la tâche la plus ardue et la plus difficile a été de transcrire l'oral. Blanche-Benveniste et Jeanjean (1987, p.93) ont observé cette difficulté d'écoute :

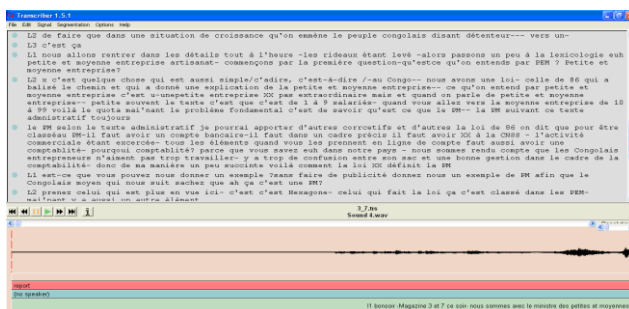
si l'on ne tient pas compte de cette grande difficulté à établir les textes de français parlé, on perpétue les pires stéréotypes ; on fournit du français parlé tel qu'on croit naïvement qu'il devrait être et non tel qu'une observation un peu exigeante le révèle.

La première difficulté est liée à « l'enregistrement ». Il y a toujours une perte des données entre l'écoute et la transcription. Une perte des sons est souvent constatée car il demeure une différence puisqu'une « écoute d'enregistrement apporte nécessairement des problèmes factices qu'on n'a

pas dans l'écoute directe. » Les difficultés d'écoute dépendent en partie de la technique utilisée lors des enregistrements.

Nos données de départ ont été enregistrées avec un dictaphone Panasonic Fast Playback/ 2- speed utilisant des micro-cassettes MC-60 d'une durée de 45 minutes. L'appareil choisi avait l'avantage d'être léger, discret et pouvait se mettre sans encombre dans la poche. Il avait aussi l'avantage de sa bonne portée. Nous avons souvent eu quelques coupures lorsqu'il s'agissait de changer de cassette. Actuellement, nous utilisons des dictaphones numériques qui ont l'avantage de ne pas utiliser de cassettes. Une fois les enregistrements réalisés, ils sont directement transférés dans l'ordinateur par le biais de logiciels spécialisés.

Comme pour les différentes transcriptions, les exigences techniques résultent, en partie, des objectifs du chercheur. Un phonéticien n'aura pas les mêmes exigences qu'un lexicographe, un ethnologue ou un syntacticien. Les transcriptions ont été réalisées par nos propres soins de deux façons. Au départ, nous transcrivions en écoutant séquence par séquence. Par la suite, nous avons utilisé directement le logiciel **Transcriber** qui offre la possibilité d'aligner du texte et du son. Les fichiers audio sont au format Wav. Ils ont été enregistrés selon les normes en la matière soit 44.1000 MHZ en 16 bits pour favoriser une bonne audition.



Blanche- Benveniste et Jeanjean (1987, p. 98) précisent que l'informateur « doit avoir une idée du but de la transcription, et doit pouvoir centrer son attention sur les aspects qu'on voudra plus particulièrement étudier ». Par la suite, nous avons procédé à une écoute attentive de nos transcriptions en vue de percevoir l'absence ou la présence de l'adverbe de négation *ne* chez certains locuteurs.

oh-toi tu es têt-tu comprends pas- (AD, 7,7)

bon si tu n'as pas l'argent tu ne les dragues pas (AD, 6,8-9)

Il y a des cas délicats où nous ne percevons pas réellement ce que dit le locuteur. Benveniste et Jeanjean citent *lequel, laquelle, lesquels* ; des constructions délicates dans l'entourage de la négation, la distinction entre *de* et *des* ; les constructions de verbes peu connues sont également difficiles à percevoir.

5.3.1. Les multi-transcriptions

Pour pallier une difficulté de choix entre plusieurs écoutes, le GARS a proposé la possibilité des multi-transcriptions. Il s'agit de la possibilité pour le transcripteur d'utiliser deux transcriptions pour des phénomènes insaisissables. Elles sont comprises entre deux barres obliques /...../. Nous avons les transcriptions de cet ex. où le clitique objet *l'/la* est attestée.

xxx l'argent on /l'a giflée, la giflait/ on la faisait n'importe quoi
(RJCE, 19,11)

que vous / mangez, mangiez /

5.3.2 Les mots incompréhensibles

Nous avons été confrontés aussi à la transcription des mots incompréhensibles dans notre corpus. Ils sont représentés dans le corpus par le X qui représente une syllabe incompréhensible tandis que XXX correspond à une suite des syllabes incompréhensibles.

hein- y a xx donc des gens de Bifouiti (C., 51, 4)

oh le pont là c'est cassé- donc vraiment que je n'avais que tout- xxx
(C., 52,1)

5. La représentativité quantitative du corpus

Le choix du corpus ouvert n'impliquait pas de parcourir tout le Congo. Le FPC est constitué des enregistrements transcrits de Brazzaville, d'autres enregistrements sont en cours de transcription. C'est pour cette raison que notre objectif n'est pas de travailler sur la représentativité quantitative du

corpus mais au contraire sur sa qualité. Tout en sachant avec Marie-Paule Jacques (2005, p27) que

la question de la représentativité est intrinsèquement liée à celle de la généralisation. La possibilité de généralisation est un autre aspect qui suscite la critique. En effet, la question se pose, pour la linguistique de corpus, de savoir jusqu'où extrapoler les observations, jusqu'où étendre les hypothèses et les explications.

Pouvons-nous donc généraliser nos observations et postuler que les restructurations observées dans notre corpus sont extrapolables à la parlure quotidienne de tous les Congolais ? Nous parlerons plutôt d'une tendance d'une généralisation dans ce sens que les restructurations sont souvent attestées hors de notre pays. Nous nous interrogeons avec Jacques pour savoir selon quels critères constituer alors un ou des corpus. Que prendre en compte pour que ces corpus soient réellement des échantillons d'emploi de la langue (et quels emplois) ? Pour Habert (2000, p. 15) :

améliorer la représentativité d'un corpus consiste à préciser la production et la réception de chacun de ses composants, en lien avec les motifs qui ont conduit à la création du corpus, mais aussi à pouvoir déterminer sur des bases objectivables les différents emplois du langage auxquels on s'intéresse.

La production de notre corpus s'appuie sur les indices linguistiques observés dans les différents textes constituant le FPC que l'on appelle *le profilage de textes*. Le FPC est constitué d'un échantillon varié de situations et des « genres » de prise de parole, de différents types d'événements de communication, en particulier les monologues, les dialogues et les conversations. Ce qui n'est pas le cas dans la constitution des particularités lexicales en français au Congo. Queffélec et Niangouna (1990, p.46) reconnaissent que

dans la pratique, l'essentiel de l'enquête a été menée de manière moins rigoureuse : les chercheurs ont utilisé des méthodes moins sophistiquées. Ils ont fait appel à la technique du questionnaire préétabli ou demandé à des interlocuteurs complaisants d'évoquer dans des

conversations plus ou moins dirigées tel ou tel thème de la vie courante.

Contrairement à ces chercheurs, notre enquête ne comportait **aucune directive à l'endroit** des interlocuteurs. Les locuteurs étaient libres d'aborder des sujets de leur choix. Ce qui explique d'ailleurs la diversité constatée dans notre corpus.

6. Problème juridique et anonymisation

La transcription des données se conforme à une déontologie, en particulier pour respecter l'anonymat des locuteurs. Elle suscite des problèmes juridiques dans la mesure où il s'agit en quelque sorte de la vie privée des locuteurs. Encrevé, cité par Blanche-Benveniste et Jeanjean (1987, p.181), va jusqu'à interdire de les publier :

Il n'est pas question de communiquer ses enregistrements. S'intéresser au langage oral, c'est s'intéresser aux gens ; transmettre une transcription de leurs enregistrements, ce serait comme passer à quelqu'un une partie de leur corps.

La publication des données a donc été soumise à des réaménagements internes. Par exemple, pour ne pas choquer les sensibilités, nous avons masqué tous les noms propres. Ils ont été remplacés par les noms forgés. Les noms des villes ou villages ont été changés. Nous pensons que, dès lors que les transcriptions ou les enregistrements ne comportent pas une atteinte à la vie privée, il est possible de les publier.

7. L'exploitation des données

Nous avons décidé d'exploiter nos données à l'aide d'une assistance informatique, c'est-à-dire aux logiciels d'ingénierie permettant le traitement du corpus et de faire des recherches précises. Dans cette section, il ne s'agit pas de faire étalage de technicité, mais juste d'exposer brièvement et simplement notre démarche et notre travail, et d'en faire ressortir la pertinence pour l'analyse de notre corpus.

L'exploitation des données a été faite simultanément avec trois logiciels. Ces logiciels présentent certaines facilités si « l'on fait appel à la fonction CONCORDANCE du menu principal, on obtient un contexte étroit qui tient en une ligne et qui montre la forme (ou l'expression) cherchée, en position centrale, avec une demi-douzaine de mots à gauche et à droite »

(Brunet, 2006, p25). Cependant les logiciels d'ingénierie n'ont pas tous les mêmes avantages.

Le premier, c'est Intex qui offre la possibilité de recenser toutes les occurrences du mot recherché. Une recherche sur *la*, nous obtenons :

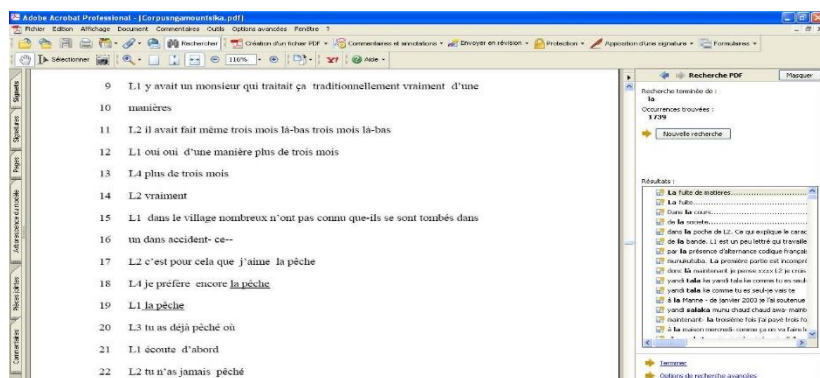
[illegible]

Le logiciel INTEX ne permet pas de distinguer le *la* clitique objet du *la* quantifiant bipolaire de même qu'il ne précise pas le texte dans lequel l'occurrence a été trouvée. Pour pallier cette absence de désambiguïsation, nous avons recouru à la dernière version du logiciel d'Hyperbase

C:\HYPERBAS\NGIM.EXE									
Forme	Lecture	Code	Syntaxe	Expr.	Initial	Final	Chain	Liste	Tout
								N° 22	CONCORDANCE
Ca	18a			xxx l' argent on / l' a gifiée , la	gifiéat / on la faisait n' importe				
Ca	18a			/ l' a gifiée la gifiéat / on la	faisait n' importe quoi alors euh q				
Ca	18a			qui va se passer au moment où on	tapait les gens s' en prenaient à c				
Co	22a			euh comment on l' appelle déjà - la	' ai oublié on a mis le nom là moi				
Co	23a			l' appelle jé l' ai remis / ta , la	/ note et il m' a même dit que c'				
Co	32a			vec réserve c' est à peine que je	regardai c' est avec réserve que				
Co	32a			gardai c' est avec réserve que je	regardai de - mes yeux néanmoins				
FU	40a			bon ils cuisaient leur nourriture	la préparait - euh et puis euh				
Ci	47a			tineraire ; puis leur étinceraire	tu fais Poto - poto ou Total est				
Ci	49a			cent la les petits cent - cent	la paient combien cinquante francs seu				
Ci	56a			peut me compléter mes salaires	on sera en égalité avec le chauffeu				
Ci	60a			arrive aussi la police . mais la	police c' est parce que vous n' avi				
Ci	65a			on (n) appelle ça quoi . euh la	patente la patente c' est pourquoi				
Gu	79a			oins à deux , à deux parcelles de	la notre . chez les ya Gayi . les garg				
Mb	133a			aïs ça se passe que son . si . si	sa . vane ça se brûle trois fois				
Mb	160a			copine qui est avec toi - mais la	copine là n' est pas fidèle bon éco				
Mb	163a			ono . moi je disais que no . donc la	/ technique , la personne / dont				
Mb	178a			pour pour maîtriser leur vie . tu	vois partout tu vois de dalles . tu				
Mb	184a			. tu sauras ah que que le petit	cravaille bien . ou bien c' est un				

Ce logiciel nous indique, d'une part, le texte dans le corpus, d'autre part, nous propose un résultat proche de nos préoccupations concernant le clitique objet *la*. Ces deux logiciels génèrent, par ailleurs, un concordancier sur le lemme sollicité. Si les données n'étaient transcrites orthographiquement, ces analyses ne seraient pas possibles. Il convient de signaler qu'une probabilité d'erreurs est toujours admise, d'où la nécessité d'une vérification manuelle.

Le troisième logiciel est adobe Acrobat qui permet aussi d'obtenir un lemme. Son usage n'est pas trop fiable. Lorsque vous aviez besoin de *la*, il recense le *la* et tous les mots contenant *la*. Il nous a servi à préciser le lemme, la page et la ligne demandés comme dans ce schéma.



L'utilisation de logiciels d'ingénierie linguistique nous entraîne directement dans la linguistique de corpus. La linguistique de corpus favorise l'utilisation de l'ordinateur pour observer la langue. Cette observation est polymorphe selon les objectifs du chercheur. En réalité, « cette notion d'observation, qu'elle soit "manuelle" ou qu'elle utilise les statistiques afin d'appropriser des quantités de données importantes, est primordiale » (Georges Williams, 2005, p.14). La linguistique de corpus est « un domaine qui s'intéresse aux textes, aux textes réels, c'est-à-dire produits pour des raisons de communication entre êtres humains et non des productions artificielles (...) » (Georges Williams, *ibid.*, 13). L'analyse informatique nous a permis d'élaborer des fréquences, des statistiques afin de quantifier les occurrences du FP. Allant dans le même sens, Dominique Willems (1998, p.84) souligne que :

L'utilisation de données langagières réelles, (...) permet (...) d'introduire dans la description quelques dimensions nouvelles : la dimension statistique, la variation et la dimension co(n)textuelle.

Nous avons choisi ces logiciels pour leur efficacité et aussi parce qu'il nous a été possible de rencontrer les concepteurs. Nous avons eu des échanges fructueux avec Max Silberztein, le concepteur d'Intex (actuellement il travaille sur Nooj) et avec Etienne Brunet (Hyperbase).

L'avantage de l'informatique, c'est surtout la vitesse, la rigueur et la représentativité des données, l'exécution des requêtes via le concordancier, face à des relevés manuels qui s'avèrent fastidieux (par exemple le comptage du déterminant *la*). Ces logiciels nous ont permis d'obtenir principalement, les occurrences brutes des déterminants, des substituts, des verbes de parole. Le recours au traitement informatique des données n'est pas à l'abri des écueils. Il est impossible à la machine par exemple de distinguer le *que* relatif du *que* subordonnant. L'analyse manuelle nous a permis de cerner au fur et à mesure des emplois concrets de ce dont nous avons besoin. Même avec Hyperbase qui comporte un bon lemmatiseur sur la recherche de *la*, le logiciel recense des occurrences non demandées. L'utilisation de l'ordinateur n'a pour but dans notre étude que de cerner les apports de cet outil dans la recherche en sciences du langage.

Conclusion :

La constitution d'un corpus est nécessaire pour mener une étude scientifique rigoureuse. Le CFPC permettra aux différents chercheurs qui s'intéressent aux questions de linguistique ou de l'oral congolais de trouver des exemples fiables et réels.

Références bibliographiques :

- Blanche-Benveniste, C. (1997). *Approches de la langue parlée en français*. Gap/Paris : Ophrys.
- Blanche-Benveniste, C. (1999). « Constitution et exploitation d'un grand corpus », *Revue française de linguistique appliquée*. (Dossier spécial Grands corpus : diversité des objectifs, variété des approches), vol., IV-1 : 65-74.

- Blanche-Benveniste, C. (2005). « L'étude grammaticale des corpus de langue parlée en français », *La Linguistique de corpus*, Presses universitaires de Rennes, Coll. 'Rivages linguistiques :47-66
- Damourette, J. et Pichon, E. (1940-1954). *Des mots à la pensée. Essai de grammaire de la langue française. Compléments*, Paris, Éditions d'Artrey & CNRS.
- Debaisieux J.-M., Deulofeu H.-J., (2012). « Une tâche à accomplir pour la linguistique française du XXIème siècle : élaborer une grammaire des usages du français » In *Langue Française*, 4,27-47
- Delcourt, C., Francard, M. & Moreau, M.-L., (1993). « Une banque de données textuelles sur la langue française en Belgique », In Latin D., Queffélec, A. Tabi-Manga, J.(éds.): *Inventaire des usages de la francophonie : nomenclatures et méthodologies*, Paris/Londres : John Libbey Eurotext, pp. 313-331.
- Fradin Bernard, *et alii.*, (2008). « Remarques sur l'usage des corpus en morphologie », *Langages*, 171,34-59
- Gadet F., (2017). « L'oralité ordinaire à l'épreuve de la mise en écrit : ce que montre la proximité », *Langages*,2017, 208,113-129
- GARRIC, N. et LONGHI, J. (2012). « L'analyse de corpus face à l'hétérogénéité des données : d'une difficulté méthodologique à une nécessité épistémologique », *Langages* 187, 3-12
- Habert B., (2000). « Des corpus représentatifs : de quoi, pour quoi, comment ? », In BILGER M. (dir.), *Linguistique sur corpus. Études et réflexions*, Perpignan, Presses Universitaires de Perpignan, pp.11-58.
- Hattiger, J.-L., (1983). *Le français populaire d'Abidjan. Un cas de pidginisation*. Abidjan : Publications de l'ILA.
- Jacques, M.-P. (2005). « Pourquoi une linguistique de corpus ? » In G., Williams (dir.) *La Linguistique de Corpus*, Presses universitaires de Rennes, 21-30.
- Kerbrat-Orecchioni, C. (1990). *Les Interactions verbales*, tome I, Paris, Armand Colin
- Klinger D. et Daniel Véronique G., (2006). « La grammaire du Français fondamental : Interrogations historiques et didactiques », *Documents pour l'histoire du français langue étrangère ou seconde* [En ligne], 36 | mis en ligne le 19 novembre 2014, consulté le 06 septembre 2017. URL:<http://dhfles.revues.org/1189>, consulté le 10 septembre 2018
- Labov, W., 1976. *Sociolinguistique*, Paris, ed. de Minuit.

- Ngamountsika, E., (2007). *Le français parlé en République du Congo : étude morphosyntaxique*, thèse de doctorat unique, Aix-en Provence 1 & Brazzaville.
- Queffélec, A., (1997). « Constitution, traitement et exploitation lexicographique », *Le corpus lexicographique. Méthodes de constitution et de gestion*, acte des troisièmes journées scientifiques du réseau thématique de recherche « Étude du français en francophonie » [dir.] Claude FREY et Daniele LATIN, pp. 353-368.
- Queffélec, A., et Niangouna, A., (1990). *Le français au Congo (RPC)*, Presse universitaire d'Aix.