Verbum XXX, 2008, 4

ÉTABLIR DES CORPUS ORAUX : CE QUE NOUS APPRENNENT LES TRANSCRIPTIONS

Paul CAPPEAU
Université de Poitiers – FoReLL

RÉSUMÉ

Les problèmes soulevés par la transcription ne concernent pas que la linguistique. D'où l'intérêt à mieux cerner les difficultés et les conséquences sur l'analyse que pose la transcription de corpus oraux.

ABSTRACT

The problems raised by the transcript do not affect linguistic. So it is important to better understand the difficulties and consequences on the analysis that the transcript of oral corpus raises.

Avec un tel titre, nul doute que la plupart des lecteurs ne passent leur chemin en pensant que le sujet les concerne peu et n'intéresse qu'une frange microscopique d'individus. Il est probablement déjà trop tard pour rattraper ces lecteurs – peu curieux –, mais disons tout de même aux audacieux qui sont arrivés à ce stade que la transcription et les difficultés d'établissement de productions orales sont des phénomènes plus fréquents qu'on ne l'imagine et qui touchent à des problèmes de compréhension bien plus vastes. Ainsi, à quelques semaines d'intervalle, deux journaux à grand tirage publiaient des informations reliées à ces aspects.

Dans un article intitulé '*Bande de traduc*', un hebdomadaire satirique écrivait le commentaire suivant :

> Enfin un peu de "doucette" dans un monde de brutes ? Vous n'y êtes pas du tout. Ces lignes sont extraites de la conférence de presse donnée, à propos de l'Afghanistan par l'inimitable général Puga, rebaptisé Père la Raclée. Mais les héros sont fatigués. Et le ministère de la Défense ayant sous-traité la transcription écrite, cela donne quelques approximations d'une grande poésie.
>
> Il fallait donc lire "**12,7**" (le calibre des mitrailleuses) et non "**doucette**", tout comme il faut comprendre missile "**sol-air**" et non "**solaire**". Aucun rapport

> avec l'indice de protection. Même si les soldats en auraient eu bien besoin. (*Le Canard Enchaîné*, septembre 2008, p. 8)

Quelques jours plus tard, *Libération* publiait un article portant sur un fait divers dans lequel on pouvait lire le passage suivant :

> Ces 59 secondes de conversation "*de mauvaise qualité audio*", une fois décryptées, vont livrer à la brigade de répression du banditisme (BRB) des indices essentiels pour identifier des membres du commando. […] Un policier de la BRB a décodé les étranges sons "*assui, achui*", que l'experte avait du mal à comprendre : "*Si, si, moi je connais, c'est l'avocat Achoui*." (*Libération*, 2 octobre 2008)

Dans les deux cas, il s'agit de difficultés d'écoute (qui ne sont pas nécessairement reconnues par l'auditeur) et qui débouchent sur des interprétations erronées. D'où l'intérêt qu'il peut y avoir pour un linguiste à se pencher de façon plus précise sur de tels phénomènes. Il peut être notamment intéressant de mieux dégager les facteurs liés au contexte qui peuvent peser sur ces difficultés de reconnaissance et engager une réflexion sur leur interprétation. Pour conduire une telle recherche, la situation de transcription (dans le but d'établir des corpus oraux), bien connue du monde des linguistes, constitue une opportunité pour recueillir des données nombreuses et variées.

1. LES ENJEUX DU PROBLÈME

Avant d'entrer dans une description des faits recueillis, il est peut-être opportun d'expliciter pourquoi ce problème est d'une grande importance dans le contexte scientifique contemporain. De nombreux projets de constitution de gros corpus ont (enfin) vu le jour en France et l'on assiste à une véritable explosion de la taille des données recueillies (Cappeau & Gadet, 2007). Cette inflation doit s'accompagner d'une réflexion sur les solutions de constitution les plus satisfaisantes (Baude, 2006 ; Bilger, 2008). Deux critères peuvent participer à l'évaluation : la lourdeur de la tâche et la qualité des données produites.

Le premier paramètre a des répercussions sur le coût (humain autant que financier) des transcriptions. Devant le travail très lourd que requiert la mise au propre de corpus oraux (Blanche-Benveniste & Jeanjean, 1986 ; Blanche-Benveniste, 1997), on peut se demander si le besoin croissant de données toujours plus nombreuse ne peut pas conduire à un infléchissement dans les pratiques de transcription. Pour l'heure plusieurs pratiques coexistent : la transcription humaine et la transcription automatique. La première suppose la formation de transcripteurs avertis, la seconde de logiciels performants. Cette dernière possibilité est souvent présentée comme la solution d'avenir. Et il est vrai qu'elle présente une rapidité assez impressionnante. Pour autant, elle semble performante plutôt pour la parole lue (comme le journal télévisé) et connaît malgré tout, par endroits, un taux d'échec encore élevé[1]. Voici un exemple[2] :

---

1 Il est souvent difficile d'obtenir des indications précises sur ce type de transcription (voir le site du Limsi par exemple).

2 Je le dois à l'amabilité de Pascal Nocera (Université d'Avignon). La transcription

(1a) il aurait pu retirer une heure une vente hors norme Cee la vente aux enchères de la collection d'André breton qui s'est achevé hier à rapporter beaucoup plus d'argent presque sourire vous conduisait vous êtes filmés (version automatique initiale)

(1b) il aurait pu en tirer une œuvre + une vente hors norme que la vente aux enchères de la collection d'André Breton qui s'est achevée hier elle a rapporté beaucoup plus d'argent que prévu + souriez vous conduisez et vous êtes filmés (version corrigée)

Lorsque l'on dispose les deux versions[3], on comprend que le statut de la version (1a) obtenue par transcription automatique doive être posé. C'est d'ailleurs vers ce point que peut se déplacer le débat (même s'il convient de l'élargir au cas de la transcription humaine) : quel état de la transcription doit être considéré comme final et donc soumis à exploitation ?

Là encore, deux pratiques peuvent être envisagées : une seule version est établie (obtenue soit par le biais d'un transcripteur humain soit par le biais d'un logiciel) et la question précédente ne se pose pas puisqu'un seul état – par force final – est produit ; deux versions différentes de la transcription sont successivement réalisées (une première selon les modalités précédentes est considérée comme une version initiale, une sorte de brouillon. Elle est, dans une deuxième étape, corrigée puis validée par un expert qui édite alors une version finale[4]). La première solution présente deux avantages : une plus grande rapidité et un coût moindre. La seconde possède les inconvénients inverses : plus faible productivité et renchérissement[5].

Le deuxième paramètre (la qualité de la transcription) peut intervenir pour départager ces deux solutions. Le corpus est en effet souvent présenté comme une ressource donnant accès à l'usage qui est fait de la langue. Il permet de disposer de données que le recours à l'intuition du locuteur ne peut produire en l'état. Pour Condamines (2005) : "le corpus est constitué pour étudier un fonctionnement linguistique particulier ou pour acquérir des connaissances". Si le corpus n'est pas seulement un recueil d'exemples

---

est due au démonstrateur LIA (transcription Speeral synchronisée). Il s'agit d'une version de démonstration effectuée en 2007 qui ne prétend pas refléter l'avancement actuel de ce type de travaux. La transcription (faite en simultané) est très rapide. Elle porte sur une édition du journal de France Inter. J'ai retenu, pour cette démonstration, un passage plus perturbé que d'autres.

3 Rappelons que l'usage est de ne pas ponctuer les productions orales et de marquer par un + les pauses. Je me place ici dans le choix d'une transcription orthographique en vue d'une exploitation syntaxique des corpus. Les mêmes questions pourraient se poser dans d'autres cadres (interactionnels, etc.) où les transcriptions doivent respecter d'autres conventions.

4 Finale est un terme très discutable puisqu'il est bien connu qu'une transcription n'est jamais réellement terminée.

5 Une solution mixte est possible : la version initiale, non validée par un expert serait mise en ligne dans l'attente que des corrections soient apportés par les utilisateurs. C'est le cas notamment du projet très intéressant Enquête Sociolinguistique à Orléans (*ESLO*).

destinés à illustrer des analyses livrées clé en main mais sert de fondement à la description, il convient que les données utilisées soient de très bonne qualité et à la mesure des ambitions déclarées.

La solution qui repose sur la formation de transcripteurs puis la vérification par un expert est certes coûteuse[6] mais elle semble la seule à même de garantir l'exigence de sérieux indispensable quand on veut travailler sur corpus. Si l'on prend l'exemple de la langue écrite, on imagine mal que les corpus soient constitués de textes produits avant l'édition (dans des versions qui seraient à l'état de brouillons) et qu'une description soit produite sans tenir compte de l'état des données utilisées. Il ne faut pas, dans le cas de l'oral, se tromper d'objectif : si l'on souhaite décrire l'usage oral des locuteurs il faut respecter aussi scrupuleusement que possible leur production. Sans cela c'est à des données approximatives, transformées, à demi attestées que l'on a affaire et l'on peut alors, à juste titre, s'interroger sur le statut de l'objet fourni.

2. TYPOLOGIE DES INTERVENTIONS

Avant de se pencher plus avant sur les erreurs de transcriptions constatées, il convient de rappeler très brièvement dans quelles conditions peut se faire le recueil de ces interventions. Deux états de la transcription sont utilisés : l'un résulte du travail d'étudiants (en général de L3) formés (c'est-à-dire avertis des difficultés, sensibilisés aux principales erreurs qu'ils sont susceptibles de commettre, familiarisés avec les conventions de transcription attendues). On la désignera comme la version initiale. L'autre, la version éditée, est celle qui est obtenue après vérification du travail précédent par un expert (c'est-à-dire un enseignant qui a déjà une expérience du travail de correction et connaît donc les principaux lieux de fragilité). Le présent travail repose donc sur la comparaison entre ces deux états[7].

De façon succincte, 5 grands types d'erreurs peuvent être identifiés[8]. Elles sont présentées dans le tableau suivant :

	Version initiale	**Version éditée**
Conventions	m'enfin	mais enfin
Oubli	vous l'acceptez vous l'acceptez pas	vous l'acceptez **ou** vous l'acceptez pas
Ajout	je **ne** me suis pas	je me suis pas
Transformation	**vous avez** la permission	**avec** la permission
Orthographe	on m'a rétabli**t** les **a**méricains	on m'a rétabli les **A**méricains

Tableau 1. *Typologie des erreurs recensées*

6 Pour l'instant, il faut reconnaître que l'aide apportée par les étudiants rend la transcription humaine moins onéreuse.

7 La version éditée peut, elle-même, résulter de plusieurs couches de correction.

8 Il n'est pas tenu compte, dans cet inventaire, de certaines rectifications comme l'ajout d'alternances auditives.

L'image suivante permet de visualiser, sur un court extrait, les interventions opérées sur la version initiale. Bien évidemment, ces changements seront ensuite intégrés dans le fichier informatique qui constituera la version éditée.

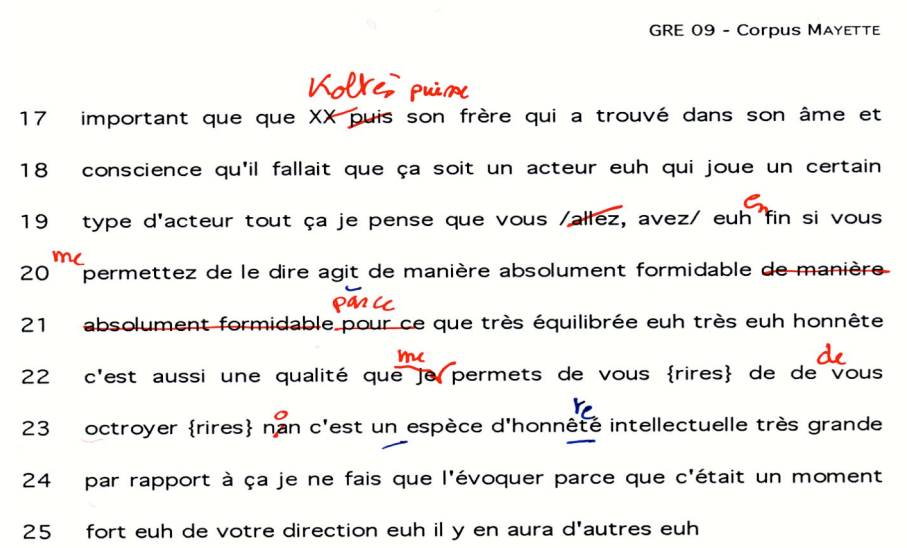

Image 1. Les modifications apportées à la version initiale

Un premier facteur, externe, peut intervenir sur la répartition de ces erreurs dans les transcriptions. En effet, selon les données orales sur lesquelles travaillent les transcripteurs, de fortes disparités apparaissent. Ainsi, pour le corpus *Phonothèque*, les étudiants étaient confrontés à des interviews d'artistes, disparus pour la plupart, évoquant des faits bien antérieurs à la naissance des transcripteurs et dont ils n'avaient probablement jamais entendu parler. De nombreuses erreurs liées à une méconnaissance culturelle peuvent alors être identifiées (mauvaise reconnaissance de noms propres ou d'œuvres), d'où l'importance des erreurs dites de transformation. Quand les transcripteurs enregistrent des personnes de leur entourage et transcrivent leur propos (corpus *jeunes locuteurs*), ces erreurs-là sont bien moindres, d'où un rééquilibrage des divers types d'erreurs (et, de façon mécanique, la part plus importante que connaissent les oublis ou les erreurs liées aux convention) :

	corpus *Phonothèque*	*Corpus jeunes locuteurs*
Conventions	8 %	27 %
Oubli	13 %	22 %
Ajout	6 %	2 %
Transformation	21 %	13 %
Orthographe	52 %	40 %

Tableau 2. *Répartition des erreurs dans deux corpus*

Sans entrer dans des indications chiffrées trop détaillées, on peut d'ailleurs remarquer que, globalement, les interventions sont moindres pour les transcriptions qui composent un corpus "proche" (comme le corpus jeunes locuteurs) que pour celles qui constituent un corpus "éloigné" (tel que *Phonothèque*).

3. HYPOTHÈSES SUR L'ORIGINE DES PROBLÈMES OBSERVÉS

Il est surprenant que la situation de transcription ne soit pas plus exploitée par les psycholinguistes qui cherchent souvent à jouer sur des contraintes multiples pour mieux comprendre certains fonctionnements cognitifs. En effet, le dépouillement des interventions entre les deux versions est particulièrement riche et permet de dégager des tendances dans les écarts des transcripteurs.

3.1. La reconnaissance des mots

Ainsi, les modèles d'identification des mots (Frauenfelder, 2002 ; Dufour & Frauenfelder, 2007) permettent de décrire certaines des transformations opérées lors des transcriptions. Très souvent, la mauvaise identification va s'accompagner de la reconstruction d'un mot ou d'une séquence qui présente des sonorités proches. Le transcripteur s'appuie de fait sur un élément sonore (phonème, syllabe) qui va servir de support à la fabrication d'une unité. Pour les exemples (2a) et (2b), la séquence initiale /lan/ constitue la partie commune et oriente le choix de la forme alternative. Le transcripteur procède à un redécoupage du syntagme (par rapport à la version finale) et propose une version compatible avec le contexte environnant :

(2a) pourquoi les écrevisses à **l'aneth** (version initiale)
(2b) pourquoi les écrevisses à **la nage** (VAL09-Cuisine – version corrigée)

Dans le cas de (3a) et (3b) la suite /iʒjø/ peut être vue comme l'élément déclencheur qui conduit le transcripteur à faire défiler son lexique mental pour s'arrêter sur le terme *religieux :*

(3a) la science va faire des progrès **religieux** + vous savez (version initiale)
(3b) la science va faire des progrès **prodigieux n'est-ce pas** vous savez (GRE09-Ormesson – version corrigée)

On peut supposer que deux facteurs favorisent le recours à l'adjectif *religieux* : d'une part, aucune collocation "progrès… /iʒjø/" n'est activée, d'autre part, la présence du terme *science* dans le voisinage antérieur proche peut avoir facilité une opposition *science* vs *religion*, qui fait partie de la doxa. Enfin, c'est toute la finale ("-ension") qui est concernée dans (4a) et (4b) et conduit, là encore, le transcripteur à produire un mot distinct de celui qui a été réalisé :

(4a) il y avait déjà une énorme **propension** de de de de ce futur maritime (version initiale)
(4b) il y avait déjà une énorme **compréhension** de de de de ce futur maritime (VAL09-Kersauson – version corrigée)

Certaines erreurs, plus rares, montrent aussi la part active du transcripteur qui procède à un remodelage de l'information perçue. D'où le recours à un synonyme en (5a) qui ne possède aucun élément sonore en commun avec la forme produite (5b) :

(5a) quand je suis revenu de ce **voyage** en Allemagne (version initiale)
(5b) quand je suis revenu de ce **séjour** en Allemagne (VAL09-Kersauson – version corrigée)

3.2. Les connaissances culturelles

La mauvaise identification d'un mot ou d'un syntagme est souvent liée à un problème de connaissance. C'est d'ailleurs pourquoi les noms propres sont souvent très perturbés : le transcripteur ne peut pas s'appuyer sur le contexte pour réexaminer / réinterpréter sa proposition. C'est le cas en (6a) et (7a) où la bonne perception de la séquence sonore n'est pas analysée comme un nom propre (6b)-(7b)[9] :

(6a) c'était une nouvelle approche la la la méthode de **marche** (version initiale)
(6b) c'était une nouvelle approche la la la méthode de **Marsch** (GRE09-Arsenic – version éditée)
(7a) **qu'est-ce qui vit** qui a dit ça sur notre plateau (version initiale)
(7b) c'est **Steevy** qui a dit ça sur notre plateau (VAL09-Allègre – version éditée)

Dès lors que le transcripteur ne possède pas une expression ou un terme dans son lexique, il ne peut plus proposer une mise en mots satisfaisante. Comme on l'a vu précédemment, il recompose une séquence à partir des éléments sonores saillants qu'il retient. Si l'on identifie bien une part active du transcripteur dans ce cas, elle est guidée par la forme sonore. Le résultat est donc souvent une parenté formelle sans aucune relation sémantique entre la séquence initiale et la version éditée. Les exemples (8)-(9) illustrent bien ce point :

(8a) ce membre distingué de l'académie française nous livre son **pince manuel** d'inspiration très autobiographique (version initiale)
(8b) ce membre distingué de l'Académie française nous livre son **pensum annuel** d'inspiration très autobiographique (GRE09-Ormesson – version éditée)
(9a) dans l'histoire **de Jules Ferrand** (version initiale)
(9b) dans l'histoire **du juif errant** GRE09-Ormesson – version éditée)

3.3. Contextes propices

L'environnement dans lequel prend place la séquence mal interprétée peut présenter certaines particularités linguistiques qui, après coup, permettent de comprendre comment l'erreur a pu être générée (Bond, 2005). Ainsi, la succession, de deux phonèmes identiques peut faciliter l'oubli de

9 Google permet, dans un tel cas, à l'expert de paraître plus savant qu'il n'est…

l'un des deux. C'est ce qui se produit par exemple pour certains clitiques. Ainsi "se" peut ne pas être perçu devant un mot qui commence par /s/, d'où l'erreur en (10) :

(10a) évidemment il ne soumettra jamais (version initiale)

(10b) évidemment il ne **se** soumettra jamais (GRE09-Onfray – version éditée)

Les bribes (i.e. réitérations d'éléments grammaticaux) sont souvent très difficiles à quantifier. On observe alors, le plus souvent, l'omission d'une occurrence dans une série (11) :

(11a) juste avant la pause (version initiale)

(11b) juste avant la la pause (GRE09-Busson – version éditée)

La situation de transcription, par ses caractéristiques particulières (fragmentation de l'écoute pour identifier un passage avec parfois focalisation sur des segments très courts) permet d'observer comment la compréhension d'un passage se met en place et notamment divers dysfonctionnements. Soumis à un flux sonore qu'il doit interpréter en dehors de la présence de l'interlocuteur, le transcripteur est conduit à "travailler" sur le matériau sonore : il prélève certains sons (phonèmes ou syllabes) qui deviennent le support de la recomposition qu'il élabore. Ces solutions alternatives révèlent aussi que le traitement de l'information reste très localisé : le transcripteur semble sensible à une cohérence très restreinte (qui peut porter sur un mot ou un syntagme) et ne donne pas l'impression, dans de tels cas, de replacer la forme active dans une séquence large.

4. CONSÉQUENCES DES ERREURS DE TRANSCRIPTION

La partie précédente proposait un examen des difficultés de transcription selon le point de vue du transcripteur. Cette partie traitera des transcriptions dans la perspective du lecteur ou de l'utilisateur de corpus. Il est temps, en effet, de s'interroger sur les conséquences que les erreurs de transcription auraient si la version initiale était rendue disponible et qu'aucune version corrigée n'était proposée (parce que le travail de révision et d'établissement des textes ne serait pas prévu et n'aurait pas été entrepris). L'incidence la plus manifeste des changements opérés de façon involontaire ou incontrôlée porte sur la qualité des données orales ainsi constituées : du point de vue du sens, elles sont généralement moins cohérentes, du point de vue syntaxique, elles peuvent devenir déconcertantes.[10]

4.1. Manque de cohérence

Pour un lecteur néophyte de corpus oraux, les transcriptions s'avèrent souvent assez complexes à lire mais, une fois certaines habitudes de lectures acquises, les textes oraux ne posent pas de gros problèmes de compréhension. Le cas de la version initiale est toutefois un peu particulier puisqu'elle

10 Les répercussions de type sociolinguistique (Gadet, 2009) ou typante (Mondada, 2002) ne seront pas discutées.

peut comporter de nombreuses erreurs qui en rendent la compréhension plus incertaine. Ainsi, une erreur d'orthographe comme en (12a) peut ralentir la compréhension en obligeant à revenir sur le contexte pour trouver la forme adaptée (12b) :

(12a) il y avait le **saut** de l'autorité politique (version initiale)

(12b) il y avait le **sceau** de l'autorité politique (GRE09-Chebel – version éditée)

Parfois, un néologisme peut être proposé (13a), ce qui là encore n'aide pas à l'interprétation du passage. Une écoute plus attentive permet de passer d'une version bancale (soulignée par le sic) à une version tout à fait banale :

(13a) simplement mille constérations faisaient que {sic} (version initiale)

(13b) simplement mille considérations faisaient que (VAL09-Bilger – version éditée)

Le remplacement d'une séquence par une autre provoque une perturbation plus étendue et peut rendre opaque un passage large. Ainsi en (14a)-(14b), on ne sait plus très bien comment relier ce "volant" à l'ensemble des acteurs qui sont énumérés (ministre, gouvernement, etc.). La difficulté de compréhension porte finalement sur la totalité de la construction :

(14a) une collaboration entre **le volant des** politiques incarnés par un ministre ou un gouvernement ou un président et une fonction publique (version initiale)

(14b) une collaboration entre **une volonté** politique incarnée par un ministre ou un gouvernement ou un président et une fonction publique (GRE09-Joxe – version éditée)

En définitive, le changement d'un mot dans un énoncé ajoute une couche de complexité à la lecture des corpus oraux. Ceux-ci sont déjà rendus difficiles à interpréter car ils contiennent de nombreuses entorses au déroulement linéaire que la langue écrite nous habitue à gérer (bribes, interruptions, parenthèses, etc.). Si, de plus, s'accumulent des syntagmes en apparence correctement formés mais sémantiquement non interprétables, on imagine sans peine les répercussions négatives (et même déplorables) sur la vision de la langue parlée qui est ainsi délivrée dans les corpus. Par contrecoup, c'est la question portant sur l'exploitation des corpus oraux (dans le but de décrire la langue) qui peut être posée.

4.2. Incertitudes syntaxiques

Ce dernier point est bien connu et a été signalé à de nombreuses reprises (Blanche-Benveniste & Jeanjean, 1986 ; Blanche-Benveniste, 1997 ; Bilger, 2008 ; Cappeau, 2008). Il n'est donc pas utile de trop le développer. On sait que la prononciation ne permet généralement pas de distinguer *qu'il* et *qui*. Le choix entre ces deux formes doit donc tenir compte du contexte car, en général, une forme est plausible syntaxiquement (15b) alors que l'autre est agrammaticale (15a) :

(15a) l'accès à la connaissance tel **qui** peut être organisé (version initiale)

(15b) l'accès à la connaissance tel **qu'il** peut être organisé (GRE09-Numérique – version éditée)

La régularité d'un passage syntaxique disparaît facilement pour peu qu'une courte séquence soit mal orthographiée et découpée. Ainsi, en (16a) l'emploi du relatif *dont* semble non maîtrisé, ce que la version (16b) contredit :

(16a) il y a un certain nombre d'éléments euh dont il euh **qu'on vient** de de tenir compte (version initiale)

(16b) il y a un certain nombre d'éléments euh dont il euh **convient** de de tenir compte (GRE09-Arsenic1 – version éditée)

La bizarrerie syntaxique du français parlé peut même être atteinte par certaines transcriptions erronées. Ainsi, une tournure surprenante (17a) "qui est en accident" ne l'est plus dans la version finale (17b) bien plus habituelle ("maquillé en") :

(17a) j'ai extrait de de votre livre euh ce ce crime **qui est** euh en accident (version initiale)

(17b) j'ai extrait de de votre livre euh ce ce crime **maquillé** euh en accident (GRE09-Surig – version éditée)

L'absence de compléments clitiques (signalé en 3.2.) peut donner l'illusion que la construction des verbes n'est pas conforme à l'usage connu, ce qui demande à être vérifié. Quant aux descriptions quantifiées sur les modes de production comme les bribes, encore faut-il que les occurrences des unités soient correctement relevées. De fait, les données non corrigées pourraient poser un vrai problème sur la nature de l'objet proposé à la description.

Là encore inutile de trop insister sur la vision du français parlé que la version initiale finirait par imposer : une série d'énoncés ininterprétables, qu'aucune description syntaxique ne parviendrait à régulariser, car ils enfreindraient systématiquement les règles d'organisation de la langue. C'est le triomphe assuré des préjugés sur l'oral qui serait ici exposé et renforcé. Le corpus pourrait aboutir au triomphe de la vision négative de l'oral dont des linguistes, par des décennies de travail, ont montré l'inanité.

5. CONCLUSION

L'attention portée à la transcription permet de rappeler les embûches de cette pratique mais rappelle aussi tout l'intérêt des faits qu'elle permet de relever. En observant la première version des transcripteurs, en s'intéressant aux écueils qu'ils doivent éviter, on dispose de données précieuses qui attestent des écarts entre ce qui est produit et ce qui est perçu. Ainsi, la transcription s'avère être un dispositif particulièrement adapté pour approcher les difficultés de compréhension que les interlocuteurs peuvent parfois éprouver

dans une situation d'échange. La transcription permet de récupérer un matériau attesté de ces difficultés en contexte.

Du côté des utilisateurs de corpus, cette réflexion permet de mesurer l'importance extrême que la phase de transcription possède : sans une vérification minutieuse, la constitution de gros corpus oraux risque d'aboutir à l'élaboration de données pseudo attestées[11]. Et cela semble encore plus net quand les transcripteurs sont soumis à des enregistrements éloignés de leurs connaissances. On voit donc aussi le risque vers l'uniformité qu'une solution à moindre coût pourrait entrainer : il suffirait de cantonner les enregistrements à des domaines maitrisés par les transcripteurs, à se contenter d'un français usuel portant sur des thèmes familiers. Il a fallu attendre de longues années pour que le regard sur le français parlé évolue, souhaitons que les corpus oraux restent bien fidèles à la langue parlé et n'en viennent pas à conserver un reflet déformé, voire informe, orientation dont les conséquences seraient ravageuses.

RÉFÉRENCES BIBLIOGRAPHIQUES

BAUDE O. (éd.) (2006), *Corpus oraux – Guide des bonnes pratiques*, Paris, CNRS Éditions.

BILGER M. (éd.) (2008), *Données orales – Les enjeux de la transcription*, Perpignan, PUP.

BLANCHE-BENVENISTE C. & JEANJEAN C. (1986), *Le français parlé. Édition et transcription*, Paris, Didier-Érudition

BLANCHE-BENVENISTE C. (1997), *Approches de la langue parlée en français*, Paris, Ophrys.

CONDAMINES A. (2005), "Sémantique et corpus, quelles rencontres possibles ?", in Condamines A. (éd.), *Sémantique et corpus*, Paris, Hermès, 15-38.

BOND Z. S. (2005), "Slips of the Ear", in Pisoni D. B. & Remez R. E. (eds), *The Handbook of Speech Perception*, Malden, Blackwell Publishing, 290-310.

CAPPEAU P. (2008), "Perception et reconstruction", in Bilger M. (éd.), *Données orales : les enjeux de la transcription*, Perpignan, PUP, 235-247.

CAPPEAU P. & GADET F. (2007), "Où en sont les corpus de français parlés ?", *Revue Française de Linguistique Appliquée*, XII-1, 129-134.

DUFOUR S. & FRAUENFELDER U. H. (2007), "L'activation et la sélection lexicale lors de la reconnaissance des mots parlés : modèles théoriques et données expérimentales", *L'année psychologique*, 1, 65-86.

FRAUENFELDER U. H. (2002), "La reconnaissance des mots parlés", in Florin A. & Morais J. (éds), *La maîtrise du langage*, Rennes, PUR, 25-39.

GADET F. (2008), "L'oreille et l'œil à l'écoute du social", in Bilger M. (éd.), *Données orales : les enjeux de la transcription*, Perpignan, PUP, 35-47.

MONDADA L. (2002), "Pratiques de transcription et effets de catégorisations", *Cahiers de praxématique,* 39, 45-75.

11 On mesure ce que cette désignation a d'absurde. C'est pourtant bien la seule adaptée aux données recueillies en version initiale dans les conditions envisagées.